Mathematics of Robustness, Risk & Failure

Volume III: Robust Optimization & Decision under Uncertainty

Minimax, distributionally robust, and adaptive frameworks for decisions when models are wrong

---

Every derivation step by step

Every symbol defined

Explicit numerical examples

Geometric intuitions

Silvano Salvador

Contents

9.1 Uncertainty Sets: Box, Ellipsoidal, and Polyhedral 3

9.2 Robust Linear and Conic Optimization 8

9.3 The Robust Counterpart and Duality 12

9.4 Price of Robustness and the Bertsimas–Sim Framework 18

9.5 Adjustable and Multi-Stage Robust Optimization 23

10.1 Ambiguity Sets: Moment-Based and ϕ-Divergence Balls 28

10.2 Wasserstein Distributionally Robust Optimization 32

10.3 Duality Results for Distributionally Robust Problems 37

10.4 Data-Driven Ambiguity Sets and Sample Complexity 42

10.5 Connections to Regularization and Adversarial Training 47

11.1 Knightian Uncertainty and Ambiguity Aversion 52

11.2 Maximin, Minimax Regret, and Hurwicz Criteria 56

11.3 Choquet Expected Utility and Non-Additive Measures 60

11.4 Robust Bayesian Decision Theory and Γ-Minimax 66

11.5 Info-Gap Decision Theory 70

12.1 Two-Stage Stochastic Programming with Recourse 74

12.2 Scenario Generation and Reduction Techniques 78

12.3 Chance-Constrained Programming 83

12.4 Multi-Stage Stochastic Programming and Decision Trees 87

12.5 Robust vs Stochastic: A Formal Comparison 91

13.1 The Core Mathematical Thread: Support Functions and Duality 97

13.2 Open Problems and Directions 98

13.3 Connections to Adjacent Fields 99

13.4 A Reader's Map of the Literature 100

A. Convex Analysis and Conic Duality Toolkit 101

B. Probabilistic Background: Concentration, Divergences, and Optimal Transport 103

C. Algorithms and Computational Complexity Reference 107

D. Notation and Glossary **109**

E. Extended Proofs and Supplementary Derivations **113**

Part 9

Foundations of Robust Optimization

The geometry of uncertainty sets and the formal calculus of worst-case optimization.

9.1 Uncertainty Sets: Box, Ellipsoidal, and Polyhedral

Overview

Robust optimization replaces a single nominal parameter vector $\mathbf{a} \in \mathbb{R}^n$ by an *uncertainty set* $\mathcal{U} \subseteq \mathbb{R}^n$, and requires feasibility/optimality for *every* $\mathbf{a} \in \mathcal{U}$. The shape of $\mathcal{U}$ encodes both the modeller's prior knowledge and the computational tractability of the resulting problem.

Example — Case study: portfolio weights under box vs. ellipsoidal return uncertainty

A two-asset portfolio has expected returns $\bar{\mathbf{a}} = (8\%, 12\%)^\top$ with uncertainty $\boldsymbol{\rho} = (2\%, 5\%)^\top$ (box) or, alternatively, an ellipsoid with $\mathbf{P} = \operatorname{diag}(2\%, 5\%)$ (so the same nominal radii, but jointly bounded). For a portfolio $\mathbf{x} = (0.6, 0.4)^\top$, the worst-case portfolio return under the box is $\bar{\mathbf{a}}^\top\mathbf{x} - \boldsymbol{\rho}^\top|\mathbf{x}| = (0.6\cdot 8 + 0.4\cdot 12) - (0.6\cdot 2 + 0.4\cdot 5) = 9.6 - 3.2 = 6.4\%$. Under the ellipsoid, the worst case is $\bar{\mathbf{a}}^\top\mathbf{x} - \|\mathbf{P}\mathbf{x}\|_2 = 9.6 - \sqrt{(0.6\cdot 2)^2 + (0.4\cdot 5)^2} = 9.6 - \sqrt{1.44 + 4} = 9.6 - \sqrt{5.44} \approx 9.6 - 2.33 = 7.27\%$. The ellipsoidal worst case (7.27%) is less pessimistic than the box worst case (6.4%) by 0.87 percentage points – in portfolio terms, this is the difference between a "both assets simultaneously underperform by their full margins" scenario (box) and a "returns move together along a single worst direction, but not independently to their individual extremes" scenario (ellipsoid), the latter being the more statistically realistic assumption whenever asset returns exhibit any common factor structure.

Box uncertainty. The simplest case, $\mathcal{U}_{\text{box}} = \{\mathbf{a} : |a_i - \bar{a}_i| \le \rho_i,\ i = 1, \dots, n\}$, treats each component independently and corresponds to an ℓ_∞-ball of radius vector $\boldsymbol{\rho}$ around the nominal value $\bar{\mathbf{a}}$. It is the loosest in correlation structure but yields linear robust counterparts.

Ellipsoidal uncertainty. $\mathcal{U}_{\text{ell}} = \{\mathbf{a} = \bar{\mathbf{a}} + \mathbf{P}\mathbf{u} : \|\mathbf{u}\|_2 \le 1\}$ for some $\mathbf{P} \in \mathbb{R}^{n\times k}$. Geometrically this is an ellipsoid centred at $\bar{\mathbf{a}}$ with shape matrix $\square = \mathbf{P}\mathbf{P}^\top$. It captures correlations between components of $\mathbf{a}$ (e.g. from a covariance estimate) and leads to second-order cone constraints.

Polyhedral uncertainty. $\mathcal{U}_{\text{poly}} = \{\mathbf{a} : \mathbf{C}\mathbf{a} \le \mathbf{d}\}$ for some matrix $\mathbf{C}$ and vector $\mathbf{d}$. This generalises the box (which is the special case $\mathbf{C} = [\mathbf{I}; -\mathbf{I}]$) and remains linear-programming friendly.

Definition: Support function

For a set $\mathcal{U} \subseteq \mathbb{R}^n$ and direction $\mathbf{x} \in \mathbb{R}^n$, the *support function* is

$$\sigma_{\mathcal{U}}(\mathbf{x}) \triangleq \sup_{\mathbf{a}\in\mathcal{U}} \mathbf{x}^\top\mathbf{a}.$$

The worst-case value of a linear functional $\mathbf{a} \mapsto \mathbf{x}^\top\mathbf{a}$ over $\mathcal{U}$ is exactly $\sigma_{\mathcal{U}}(\mathbf{x})$, and this is the central object that the robust counterpart must compute or bound.

Example — Case study: a robust blending problem

A refinery blends two crude oils ($x_1, x_2 \geq 0$ barrels) to meet a minimum octane requirement $a_1 x_1 + a_2 x_2 \geq 90$, where (a_1, a_2) – octane ratings – are uncertain in a box around $(87, 93)$ with $\rho = (1, 1)$. Minimizing cost $c_1 x_1 + c_2 x_2$ with $(c_1, c_2) = (70, 75)$ (\$/barrel) subject to a $\geq$-type RC requires the *worst case to be the binding direction*: for a $\geq$ constraint, the RC is $\min_{\mathbf{a} \in \mathcal{U}} \mathbf{a}^\top \mathbf{x} \geq 90$, i.e. $(\bar{\mathbf{a}} - \boldsymbol{\rho})^\top \mathbf{x} \geq 90 \iff 86 x_1 + 92 x_2 \geq 90$. Compare the nominal-only solution $x_2 = 90/93 \approx 0.968$ (with $x_1 = 0$, cost ≈ 72.6) against the robust solution $x_2 = 90/92 \approx 0.978$ (cost ≈ 73.4) – an extra $\approx \$0.8$/barrel, the explicit price of robustness (Section 9.4) for this 1-point octane uncertainty, computed here directly from the sign-flipped support function $\sigma_{\mathcal{U}}(-\mathbf{x}) = -(\bar{\mathbf{a}} - \boldsymbol{\rho})^\top \mathbf{x}$ for $\mathbf{x} \geq 0$.

Theorem (closed-form support functions). For the three uncertainty sets above,

$$\sigma_{\mathcal{U}_{\text{box}}}(\mathbf{x}) = \bar{\mathbf{a}}^\top \mathbf{x} + \boldsymbol{\rho}^\top |\mathbf{x}|, \tag{1}$$

$$\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) = \bar{\mathbf{a}}^\top \mathbf{x} + \left\| \mathbf{P}^\top \mathbf{x} \right\|_2, \tag{2}$$

$$\sigma_{\mathcal{U}_{\text{poly}}}(\mathbf{x}) = \begin{cases} \min_{\mathbf{y} \geq 0} \mathbf{d}^\top \mathbf{y} \text{ s.t. } \mathbf{C}^\top \mathbf{y} = \mathbf{x} & \text{if feasible} \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

Proof sketch. Eq. (1) follows since $\sup_{|u_i| \leq \rho_i} \sum_i x_i (\bar{a}_i + u_i) = \bar{\mathbf{a}}^\top \mathbf{x} + \sum_i \rho_i |x_i|$, attained at $u_i = \rho_i \operatorname{sgn}(x_i)$. Eq. (2) follows from Cauchy–Schwarz: $\sup_{\|\mathbf{u}\| \leq 1} \mathbf{x}^\top (\bar{\mathbf{a}} + \mathbf{P}\mathbf{u}) = \bar{\mathbf{a}}^\top \mathbf{x} + \sup_{\|\mathbf{u}\| \leq 1} (\mathbf{P}^\top \mathbf{x})^\top \mathbf{u} = \bar{\mathbf{a}}^\top \mathbf{x} + \left\| \mathbf{P}^\top \mathbf{x} \right\|_2$. Eq. (3) is the LP dual of $\sup\{\mathbf{x}^\top \mathbf{a} : \mathbf{C}\mathbf{a} \leq \mathbf{d}\}$. ■

Example — Numerical illustration in $\mathbb{R}^2$

Let $\bar{\mathbf{a}} = (2, 3)^\top$, $\boldsymbol{\rho} = (0.5, 0.2)^\top$, and $\mathbf{x} = (1, -4)^\top$. Then $\sigma_{\mathcal{U}_{\text{box}}}(\mathbf{x}) = 2(1) + 3(-4) + (0.5 \cdot |1| + 0.2 \cdot |-4|) = -10 + (0.5 + 0.8) = -8.7$. For an ellipsoid with $\mathbf{P} = \operatorname{diag}(0.5, 0.2)$ (same nominal radii, but coupled via the ℓ_2 norm), $\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) = -10 + \left\| (0.5, -0.8)^\top \right\|_2 = -10 + \sqrt{0.25 + 0.64} = -10 + 0.943 = -9.057$. The ellipsoidal bound is tighter (less pessimistic) for this $\mathbf{x}$ because it does not allow the simultaneous worst case in *every* coordinate, only along directions consistent with $\|\mathbf{u}\|_2 \leq 1$.

General norm-ball uncertainty. The box and ellipsoid are special cases of $\mathcal{U}_p = \{\bar{\mathbf{a}} + \mathbf{P}\mathbf{u} : \|\mathbf{u}\|_p \leq 1\}$ for $p \in [1, \infty]$: $p = \infty$ gives (an affinely transformed) box, $p = 2$ gives the ellipsoid, and $p = 1$ gives a (transformed) cross-polytope. By the same Hölder-duality argument as Eq. (9.1.ell),

$$\sigma_{\mathcal{U}_p}(\mathbf{x}) = \bar{\mathbf{a}}^\top \mathbf{x} + \left\| \mathbf{P}^\top \mathbf{x} \right\|_{p^*}, \qquad \frac{1}{p} + \frac{1}{p^*} = 1,$$

so $p = 1$ (cross-polytope uncertainty) yields $\left\| \mathbf{P}^\top \mathbf{x} \right\|_\infty$ in the robust counterpart, a finite (but generally exponential in the naive formulation, polynomial via auxiliary variables) set of linear constraints. This p-p^* duality is the single most-reused identity in this entire volume.

Example — Intersection of uncertainty sets

A modeller may have *both* a box bound (componentwise sensor tolerances ρ_i) *and* an ellipsoidal bound (a covariance-based confidence region) on the same uncertain vector $\mathbf{a}$. The natural combined uncertainty set is $\mathcal{U} = \mathcal{U}_{\text{box}} \cap \mathcal{U}_{\text{ell}}$. Its support function satisfies $\sigma_{\mathcal{U}}(\mathbf{x}) \leq \min\left(\sigma_{\mathcal{U}_{\text{box}}}(\mathbf{x}), \sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x})\right)$ (intersection shrinks the set, hence the support function), but *equality need not hold* – the support function of an intersection is

in general the *infimal convolution* of the two support functions' conjugates, not their pointwise minimum. For the numerical instance of the worked example above ($\bar{\mathbf{a}} = (2,3)^\top, \mathbf{x} = (1,-4)^\top$), $\sigma_{\mathcal{U}_{\text{box}}}(\mathbf{x}) = -8.7$ and $\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) = -9.057$, so $\sigma_{\mathcal{U}_{\text{box}} \cap \mathcal{U}_{\text{ell}}}(\mathbf{x}) \le -9.057$; computing the exact value requires solving a small convex program, but the bound -9.057 is itself already a valid (if slightly loose) robust constraint, often used in practice for its closed form.

Numerical verification of the support-function identities. It is instructive to verify Eq. (9.1.box) by brute-force enumeration in a small case: $n = 2$, $\bar{\mathbf{a}} = (0,0)^\top$, $\boldsymbol{\rho} = (1,1)^\top$, $\mathbf{x} = (1,1)^\top$. The box $\mathcal{U}_{\text{box}} = [-1,1]^2$ has four corners $(\pm 1, \pm 1)$, giving $\mathbf{x}^\top \mathbf{a} \in \{2, 0, 0, -2\}$; the maximum is 2. The formula gives $\sigma(\mathbf{x}) = 0 + (1 \cdot |1| + 1 \cdot |1|) = 2$. ✓. For the ellipsoid $\mathcal{U}_{\text{ell}} = \{\|\mathbf{u}\|_2 \le 1\}$ (the unit disc, $\mathbf{P} = \mathbf{I}$), the formula gives $\sigma(\mathbf{x}) = \|(1,1)^\top\|_2 = \sqrt{2} \approx 1.414$, attained at $\mathbf{u}^* = \mathbf{x}/\|\mathbf{x}\|_2 = (1/\sqrt{2}, 1/\sqrt{2})^\top$, on the boundary of the disc in the direction of $\mathbf{x}$ – as expected, $\sqrt{2} < 2$: the disc is strictly contained in the square $[-1,1]^2$ (touching it only at the axis midpoints, not the corners), so its support function is uniformly $\le$ that of the box in every direction, with equality only along the axes.

Geometric intuition and connections. The three uncertainty sets form a hierarchy of conservativeness for fixed "nominal radius": box $\supseteq$ ellipsoid $\supseteq$ (typically) polyhedral cuts of the box. Choosing $\mathcal{U}$ is itself a modelling decision trading off tractability (box and polyhedral give LPs; ellipsoidal gives SOCPs), conservativeness (larger $\mathcal{U} \Rightarrow$ more pessimistic, but more protected solutions), and statistical fidelity (ellipsoids naturally encode covariance structure estimated from data, as developed further in Section 9.4 and Part 10).

Theorem: support function of a Minkowski sum and difference of uncertainty sets

For uncertainty sets $\mathcal{U}_1, \mathcal{U}_2 \subseteq \mathbb{R}^n$, the Minkowski sum $\mathcal{U}_1 \oplus \mathcal{U}_2 = \{\mathbf{a}_1 + \mathbf{a}_2 : \mathbf{a}_i \in \mathcal{U}_i\}$ satisfies

$$\sigma_{\mathcal{U}_1 \oplus \mathcal{U}_2}(\mathbf{x}) = \sigma_{\mathcal{U}_1}(\mathbf{x}) + \sigma_{\mathcal{U}_2}(\mathbf{x}).$$

Proof. $\sup_{\mathbf{a}_1 \in \mathcal{U}_1, \mathbf{a}_2 \in \mathcal{U}_2} \mathbf{x}^\top(\mathbf{a}_1 + \mathbf{a}_2) = \sup_{\mathbf{a}_1 \in \mathcal{U}_1} \mathbf{x}^\top \mathbf{a}_1 + \sup_{\mathbf{a}_2 \in \mathcal{U}_2} \mathbf{x}^\top \mathbf{a}_2$, since the two suprema are over independent variables and the objective separates additively. ■

Example — Compounding two independent uncertainty sources

A measured quantity $\mathbf{a} = \mathbf{a}_{\text{sensor}} + \mathbf{a}_{\text{model}}$ combines sensor noise $\mathbf{a}_{\text{sensor}} \in \mathcal{U}_{\text{box}}(\boldsymbol{\rho} = (0.1, 0.1))$ and model-discretization error $\mathbf{a}_{\text{model}} \in \mathcal{U}_{\text{ell}}(\mathbf{P} = 0.2\mathbf{I})$. For $\mathbf{x} = (1,1)^\top$: $\sigma_{\mathcal{U}_{\text{box}}}(\mathbf{x}) = 0.1 + 0.1 = 0.2$ (centred at $\mathbf{0}$), $\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) = 0.2\|(1,1)^\top\|_2 = 0.2\sqrt{2} \approx 0.283$. By the theorem, $\sigma_{\mathcal{U}_{\text{box}} \oplus \mathcal{U}_{\text{ell}}}(\mathbf{x}) \approx 0.2 + 0.283 = 0.483$. Crucially, this additive combination is *exact* – no approximation – and requires only the two individual support functions, never the explicit (and generally complicated) shape of $\mathcal{U}_{\text{box}} \oplus \mathcal{U}_{\text{ell}}$ itself (a "rounded box"). This additivity is the uncertainty-set-level analogue of the additive combination of RO and SP penalties noted in Section 12.5's worked example, and is the standard way multiple independent error sources are combined in robust engineering design (e.g. root-sum-of-squares tolerance stacking is the ℓ_2/ellipsoidal special case).

Remarks and extensions. The uncertainty-set taxonomy of this section extends naturally to *conic* uncertainty: replacing the Euclidean ball $\{\|\mathbf{u}\|_2 \le 1\}$ in the ellipsoidal definition by a general norm ball $\{\|\mathbf{u}\| \le 1\}$ for an arbitrary norm $\|\cdot\|$ yields $\sigma_{\mathcal{U}}(\mathbf{x}) = \bar{\mathbf{a}}^\top \mathbf{x} + \|\mathbf{P}^\top \mathbf{x}\|_*$ where $\|\cdot\|_*$ is the *dual norm*. Every uncertainty set considered in this textbook is, in fact, an instance of this single template for an appropriate choice of norm: $\|\cdot\|_\infty^* = \|\cdot\|_1$ recovers the box, $\|\cdot\|_2^* =$

$\|\cdot\|_2$ (self-dual) recovers the ellipsoid. Readers familiar with the dual-norm calculus from the optimization-landscape volumes of the companion Mathematics of Neural Networks series will recognise this as the same Hölder-duality machinery used there to bound gradient steps and Lipschitz constants – robust optimization and optimization-landscape analysis share a common convex-analytic vocabulary.

A further practical remark concerns *data-driven* construction of $\mathcal{U}$: given historical observations $\{\mathbf{a}^{(1)}, \dots, \mathbf{a}^{(N)}\}$ of the uncertain parameter, a natural ellipsoidal $\mathcal{U}$ is the $(1-\delta)$-confidence ellipsoid of the empirical mean and covariance, $\mathbf{P} = \hat{\Box}^{1/2}\sqrt{\chi^2_{n,1-\delta}}$ (using the χ^2 quantile for an n-dimensional Gaussian approximation), directly connecting the abstract geometric uncertainty set to a concrete statistical procedure – the data-driven analogue of the moment-based ambiguity sets that will be developed more fully in Section 10.1.

Theorem: uncertainty-set projections and marginal robustness

If $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2) \in \mathcal{U} \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and $\mathbf{x} = (\mathbf{x}_1, \mathbf{0})$ (only the first block of variables enters the objective), then $\sigma_{\mathcal{U}}(\mathbf{x}) = \sigma_{\Pi_1(\mathcal{U})}(\mathbf{x}_1)$, where Π_1 is projection onto the $\mathbf{a}_1$-coordinates – robustness against $\mathcal{U}$ for a partial-coordinate objective depends only on $\mathcal{U}$'s *projection* (marginal uncertainty set), not its full joint structure.

Proof. $\sup_{(\mathbf{a}_1,\mathbf{a}_2)\in\mathcal{U}} \mathbf{x}_1^\top \mathbf{a}_1 = \sup_{\mathbf{a}_1\in\Pi_1(\mathcal{U})} \mathbf{x}_1^\top \mathbf{a}_1$ since the objective does not depend on $\mathbf{a}_2$, so the supremum over $\mathcal{U}$ collapses to a supremum over its projection. ■

Example — Implication for correlated multi-row uncertainty

If $\mathcal{U} \subseteq \mathbb{R}^4$ jointly bounds $(a_{11}, a_{12}, a_{21}, a_{22})$ (two constraint rows, two coefficients each) via a single ellipsoid with off-diagonal correlation terms, but constraint 1's RC depends only on (a_{11}, a_{12}), then by the theorem the RC uses $\sigma_{\Pi_{\{1,2\}}(\mathcal{U})}$ – the projection of the 4-D ellipsoid onto its first two coordinates, itself an ellipsoid (projections of ellipsoids are ellipsoids) with shape matrix equal to the corresponding 2×2 principal submatrix of the original 4×4 shape matrix. **Caution**: this means that, even though the *joint* uncertainty set encodes correlation between row 1's and row 2's coefficients, each row's *individual* RC (computed via projection, as is standard practice) does *not* use this cross-row correlation – exactly the per-row-decomposition conservatism flagged in Section 9.2's remarks, now traced to its root cause: per-row RCs use marginal (projected) uncertainty sets, discarding joint/correlation information that a fully joint RC (Section 9.2's remarks, the λ_i-weighted combined-direction approach) would retain.

Historical notes. The systematic study of uncertainty sets as a modelling primitive dates to Soyster (1973), who first proposed the box-uncertainty robust LP – though Soyster's formulation was recognised as overly conservative (it corresponds to $\Gamma_i = n_i$ in the later Bertsimas–Sim notation of Section 9.4) and saw limited practical adoption for two decades. The ellipsoidal formulation, and the broader recognition that uncertainty-set *geometry* determines the robust counterpart's complexity class, is due primarily to Ben-Tal and Nemirovski (1998, 1999) and, independently, El Ghaoui and Lebret (1997) – the latter motivated by robust least-squares problems in signal processing, an application area that remains a major consumer of ellipsoidal-uncertainty methods. The general norm-ball/dual-norm formulation presented in this section's Remarks synthesises these strands and is now standard textbook material (Ben-Tal, El Ghaoui, Nemirovski, *Robust Optimization*, 2009), the primary reference for Part 9 as a whole.

Remark: rotated box uncertainty is generally not a box, and loses coordinate-wise separability

For a box uncertainty set $\mathcal{U} = \{\mathbf{a} : |a_i - \bar{a}_i| \leq \rho_i,\ i = 1, \ldots, n\}$ and an orthogonal rotation $\mathbf{R}$ ($\mathbf{R}^\top\mathbf{R} = \mathbf{I}$), the rotated set $\mathbf{R}\mathcal{U} = \{\mathbf{R}\mathbf{a} : \mathbf{a} \in \mathcal{U}\}$ is, for generic $\mathbf{R}$, *not* itself a box in the standard coordinate axes – it is a (possibly rotated) parallelepiped whose support function, by the scaling theorem of this section, is $\sigma_{\mathbf{R}\mathcal{U}}(\mathbf{x}) = \sigma_{\mathcal{U}}(\mathbf{R}^\top\mathbf{x}) = \bar{\mathbf{a}}^\top(\mathbf{R}^\top\mathbf{x}) + \boldsymbol{\rho}^\top|\mathbf{R}^\top\mathbf{x}|$, a sum of absolute values of *linear combinations* of $\mathbf{x}$'s coordinates rather than of the coordinates themselves.

Why this matters practically. The whole appeal of box uncertainty (Section 9.1's opening discussion) is that its robust counterpart $\sigma_{\mathcal{U}}(\mathbf{x}) = \bar{\mathbf{a}}^\top\mathbf{x} + \boldsymbol{\rho}^\top|\mathbf{x}|$ is coordinate-wise separable – each $|x_i|$ term depends only on the i-th coordinate, making the RC's LP reformulation (via $|x_i| \leq t_i$ auxiliary variables, one pair of linear constraints per coordinate) especially simple and sparse. Once the uncertainty is expressed in a rotated frame – which happens naturally whenever the uncertain parameters are, say, principal components of a covariance matrix rather than the original physical coordinates – this separability is lost: $|(\mathbf{R}^\top\mathbf{x})_i| = \left|\sum_j R_{ji}x_j\right|$ mixes every coordinate of $\mathbf{x}$ together, and the auxiliary-variable LP reformulation, while still exact and still tractable (the resulting RC is still an LP, just a denser one), no longer has the sparse, decoupled structure that made box uncertainty attractive in the first place.

Example — A worked numerical illustration

For $\bar{\mathbf{a}} = (2, 3)^\top$, $\boldsymbol{\rho} = (0.5, 1)^\top$, $\mathbf{x} = (1, -2)^\top$, and $\mathbf{R}$ a 45° rotation: direct computation gives $\sigma_{\mathcal{U}}(\mathbf{x}) = \bar{\mathbf{a}}^\top\mathbf{x} + \boldsymbol{\rho}^\top|\mathbf{x}| = (2 - 6) + (0.5 + 2) = -4 + 2.5 = -1.5$ for the *unrotated* box, but $\sigma_{\mathbf{R}\mathcal{U}}(\mathbf{x}) = \bar{\mathbf{a}}^\top(\mathbf{R}^\top\mathbf{x}) + \boldsymbol{\rho}^\top|\mathbf{R}^\top\mathbf{x}| \approx -5.303$ for the rotated version – a substantially different value (here, more negative, i.e. less binding as an upper bound on the corresponding constraint's left-hand side) for the *same* nominal box half-widths $\boldsymbol{\rho}$, purely as a consequence of the coordinate frame in which the box is expressed. A practitioner who fits an uncertainty box to data expressed in principal-component coordinates, then naively reapplies the box formula in the original physical coordinates (or vice versa), will compute the wrong robust counterpart entirely – the rotation must be tracked explicitly via the $\mathbf{R}^\top\mathbf{x}$ substitution, not absorbed into a simple rescaling of $\boldsymbol{\rho}$.

Connection to ellipsoidal uncertainty. This rotation-sensitivity is precisely why ellipsoidal uncertainty sets (Section 9.1's next construction) are often preferred whenever the natural uncertainty geometry is not aligned with the coordinate axes: an ellipsoid $\{\mathbf{a} : \mathbf{a} = \bar{\mathbf{a}} + \mathbf{P}\mathbf{u}, \|\mathbf{u}\|_2 \leq 1\}$ is rotation-*covariant* in a much cleaner sense, since $\mathbf{R}\{\bar{\mathbf{a}} + \mathbf{P}\mathbf{u}\} = \mathbf{R}\bar{\mathbf{a}} + (\mathbf{R}\mathbf{P})\mathbf{u}$ is again an ellipsoid of the identical shape (just recentred and with its shape matrix correspondingly rotated), and its support function's $\left\|\mathbf{P}^\top\mathbf{x}\right\|_2$ term is rotation-*invariant* when $\mathbf{x}$ and the ellipsoid are rotated together – a structural robustness-to-coordinate-choice that box uncertainty's coordinate-wise separability, for all its computational convenience, does not share.

Theorem: scaling and affine invariance of support functions

For any invertible $\mathbf{T} \in \mathbb{R}^{n\times n}$ and the transformed uncertainty set $\mathbf{T}\mathcal{U} = \{\mathbf{T}\mathbf{a} : \mathbf{a} \in \mathcal{U}\}$,

$$\sigma_{\mathbf{T}\mathcal{U}}(\mathbf{x}) = \sigma_{\mathcal{U}}(\mathbf{T}^\top\mathbf{x}).$$

Proof. $\sup_{\mathbf{a}\in\mathcal{U}} \mathbf{x}^\top(\mathbf{T}\mathbf{a}) = \sup_{\mathbf{a}\in\mathcal{U}} (\mathbf{T}^\top\mathbf{x})^\top\mathbf{a} = \sigma_{\mathcal{U}}(\mathbf{T}^\top\mathbf{x})$. ■

Example — Unit-conversion invariance of robust constraints

A robust constraint calibrated in metric units ($\mathcal{U} \subset \mathbb{R}^2$ for, say, kilograms and metres) is re-expressed in imperial units via $\mathbf{T} = \mathrm{diag}(2.205, 3.281)$ (kg→lb, m→ft). The theorem guarantees $\sigma_{\mathbf{T}\mathcal{U}}(\mathbf{x}) = \sigma_{\mathcal{U}}(\mathbf{T}^\top \mathbf{x})$, so a practitioner working in imperial units can either (i) transform $\mathcal{U}$ itself and recompute, or (ii) keep $\mathcal{U}$ in metric form and simply pre-multiply $\mathbf{x}$ by $\mathbf{T}^\top$ before evaluating the original support function – the two routes are guaranteed to agree exactly, a useful consistency check when robust models are assembled from sub-models built by different teams in different unit conventions.

Remarks. This invariance underlies a frequently overlooked pitfall in applied robust optimization: an uncertainty set calibrated from data in one coordinate system (e.g. standardized/whitened features in a machine-learning pipeline) does *not* automatically transfer its radius parameters unchanged to a re-parametrized version of the same problem. Practitioners re-deriving $\mathcal{U}$ after a change of variables (e.g. log-transforming a strictly positive uncertain parameter) must apply the theorem's $\mathbf{T}^\top$ map explicitly to the dual/direction vector, not merely rescale the radius heuristically – a common source of silently mis-calibrated robust models in practice.

9.2 Robust Linear and Conic Optimization

Overview

A robust linear program (RLP) requires $\mathbf{a}_i^\top \mathbf{x} \le b_i$ to hold for every $\mathbf{a}_i \in \mathcal{U}_i$, $i = 1, \dots, m$. The *robust counterpart* (RC) is the deterministic optimization problem obtained by replacing each uncertain constraint with its worst-case version $\sigma_{\mathcal{U}_i}(\mathbf{x}) \le b_i$, using the support functions of Section 9.1.

Example — Case study: duality certificate for a portfolio VaR-type constraint

A constraint $\mathbf{a}^\top \mathbf{x} \le b$ with $\mathbf{a} \in \mathcal{U} = \{\bar{\mathbf{a}} + \mathbf{P}\mathbf{u} : \|\mathbf{u}\|_\infty \le 1\}$ (a box after the linear map $\mathbf{P}$) corresponds, by the $p = \infty, p^* = 1$ case of the general norm-ball identity, to $\sigma_{\mathcal{U}}(\mathbf{x}) = \bar{\mathbf{a}}^\top \mathbf{x} + \|\mathbf{P}^\top \mathbf{x}\|_1$. By Theorem 9.3's dual, $\|\mathbf{P}^\top \mathbf{x}\|_1 = \sigma_{\{\|\mathbf{u}\|_\infty \le 1\}}(\mathbf{P}^\top \mathbf{x}) = \min_{\boldsymbol{\lambda}} \{\mathbf{1}^\top \boldsymbol{\lambda} : -\boldsymbol{\lambda} \le \mathbf{P}^\top \mathbf{x} \le \boldsymbol{\lambda}\}$ – i.e. $\boldsymbol{\lambda} = |\mathbf{P}^\top \mathbf{x}|$ componentwise, with $\mathbf{1}^\top \boldsymbol{\lambda} = \sum_j |(\mathbf{P}^\top \mathbf{x})_j|$, recovering $\|\mathbf{P}^\top \mathbf{x}\|_1$ directly. For a numerical instance $\mathbf{P}^\top \mathbf{x} = (3, -2, 1)^\top$: dual variables $\boldsymbol{\lambda} = (3, 2, 1)^\top \ge |\mathbf{P}^\top \mathbf{x}|$ (componentwise, tight), dual objective $\mathbf{1}^\top \boldsymbol{\lambda} = 6 = \|(3, -2, 1)^\top\|_1$ ✓– the dual LP's optimal value reproduces the ℓ_1 norm exactly, as the theorem guarantees.

Robust counterpart of an LP. Consider the nominal LP $\min_{\mathbf{x}} \mathbf{c}^\top \mathbf{x}$ subject to $\mathbf{a}_i^\top \mathbf{x} \le b_i$, $i = 1, \dots, m$, $\mathbf{x} \ge 0$, where each row $\mathbf{a}_i$ is uncertain, $\mathbf{a}_i \in \mathcal{U}_i$. The RC is

$$\min_{\mathbf{x} \ge 0} \ \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \sigma_{\mathcal{U}_i}(\mathbf{x}) \le b_i, \quad i = 1, \dots, m.$$

Theorem (RC of a box-uncertain LP is an LP). If $\mathcal{U}_i = \{\mathbf{a} : |a_{ij} - \bar{a}_{ij}| \le \rho_{ij}\}$, then using Eq. (9.1.box) the RC becomes

$$\min_{\mathbf{x} \ge 0} \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \bar{\mathbf{a}}_i^\top \mathbf{x} + \boldsymbol{\rho}_i^\top |\mathbf{x}| \le b_i.$$

Since $\mathbf{x} \ge 0$ implies $|\mathbf{x}| = \mathbf{x}$, this reduces to $(\bar{\mathbf{a}}_i + \boldsymbol{\rho}_i)^\top \mathbf{x} \le b_i$: *the box-robust LP is simply the nominal LP evaluated at the worst-case coefficients $\bar{a}_{ij} + \rho_{ij}$ in every row*, still an LP with the same number of variables and constraints.

Theorem: RC of an ellipsoidal-uncertain LP is an SOCP

If $\mathcal{U}_i = \{\bar{\mathbf{a}}_i + \mathbf{P}_i\mathbf{u} : \|\mathbf{u}\|_2 \le 1\}$, then by Eq. (9.1.ell) constraint i of the RC is

$$\bar{\mathbf{a}}_i^\top \mathbf{x} + \|\mathbf{P}_i^\top \mathbf{x}\|_2 \le b_i,$$

a second-order cone (SOC) constraint in $\mathbf{x}$. The RC of the whole LP is therefore a *second-order cone program* (SOCP), solvable in polynomial time by interior-point methods.

Example — Case study: supply chain with budgeted supplier disruptions

A company sources a component from $n_i = 6$ suppliers, each contributing $\hat{a}_{ij}|x_j| =$ (in \$1000 units) $(12, 10, 8, 6, 4, 2)$ to a cost-overrun constraint (sorted descending). Setting $\Gamma_i = 2$ ("at most 2 suppliers disrupted simultaneously," a managerial judgment based on historical disruption co-occurrence) gives worst-case overrun $\beta_i(2) = 12 + 10 = 22$ (\$1000). Compare $\Gamma_i = 0$ (no protection, $\beta_i = 0$), $\Gamma_i = 1$ ($\beta_i = 12$), $\Gamma_i = 3$ ($\beta_i = 12{+}10{+}8 = 30$), $\Gamma_i = 6$ ($\beta_i = 42$, full box). The Chernoff bound $\exp(-\Gamma_i^2/2n_i)$ for $\Gamma_i = 2, n_i = 6$ gives $\exp(-4/12) = \exp(-0.333) \approx 0.717$ – a fairly weak (71.7%) bound on the violation probability for this small n_i, illustrating that the probabilistic guarantee of Theorem 9.4 is most informative for *large* n_i (many independent suppliers); for $n_i = 6$, the choice of Γ_i is better guided by the direct cost trade-off (22 vs. 12 vs. 30 vs. 42, in \$1000) than by the (here, loose) Chernoff bound.

Example — A two-variable robust LP

Minimize $x_1 + x_2$ subject to $\mathbf{x} \ge 0$ and $a_1x_1 + a_2x_2 \le 10$, where (a_1, a_2) lies in the ellipsoid centred at $(2, 3)$ with $\mathbf{P} = \mathrm{diag}(1, 1)$ (i.e. a disc of radius 1). The RC constraint is $2x_1 + 3x_2 + \sqrt{x_1^2 + x_2^2} \le 10$. At the candidate point $\mathbf{x} = (2, 2)$: $2(2) + 3(2) + \sqrt{4+4} = 4 + 6 + 2.828 = 12.828 > 10$, infeasible. At $\mathbf{x} = (1, 1.5)$: $2 + 4.5 + \sqrt{1 + 2.25} = 6.5 + 1.803 = 8.303 \le 10$, feasible, with objective 2.5. Note that the nominal-only constraint $2x_1 + 3x_2 \le 10$ would have permitted $\mathbf{x} = (1, 2.67)$ with objective 3.67 – robustness shrinks the feasible region and raises the achievable minimum, a quantitative "price" formalised in Section 9.4.

Robust conic programs. The RC machinery extends beyond LPs to general conic programs $\min\{\mathbf{c}^\top\mathbf{x} : \mathbf{A}\mathbf{x} \preceq_K \mathbf{b}\}$ for a closed convex cone K (e.g. $K = \mathbb{R}^m_{\ge 0}$ for LP, $K =$ SOC for SOCP, $K =$ PSD cone for SDP). If row i of $(\mathbf{A}, \mathbf{b})$ is uncertain in $\mathcal{U}_i$, the RC constraint becomes $\sigma_{\mathcal{U}_i}(\mathbf{x}; K_i^*) \le 0$ where $\sigma_{\mathcal{U}_i}(\cdot; K_i^*)$ is the support function evaluated against the relevant dual-cone functional – the same machinery as Eq. (9.1.box)–(9.1.poly), but now K-valued rather than scalar.

Example — A robust SDP constraint (numerical sketch)

Consider the PSD constraint $\mathbf{A}_0 + \sum_j x_j\mathbf{A}_j \succeq 0$ where $\mathbf{A}_1$ is uncertain, $\mathbf{A}_1 = \bar{\mathbf{A}}_1 + \Delta$, $\|\Delta\|_F \le \rho$ (Frobenius-norm ball). The robust constraint $\inf_{\|\Delta\|_F \le \rho} [\mathbf{A}_0 + x_1(\bar{\mathbf{A}}_1 + \Delta) + \sum_{j\ge 2} x_j\mathbf{A}_j] \succeq 0$ can be shown to be equivalent to $\mathbf{A}_0 + x_1\bar{\mathbf{A}}_1 + \sum_{j\ge 2} x_j\mathbf{A}_j \succeq |x_1|\rho\,\mathbf{I}$ (worst-case Δ shifts the spectrum down uniformly by $|x_1|\rho$), itself an LMI in $\mathbf{x}$ after introducing the auxiliary scalar $|x_1|$ via two linear inequalities $-t \le x_1 \le t$. E.g. with 2×2 matrices, $\bar{\mathbf{A}}_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $\rho = 0.3$, $x_1 = 2$: the robust LMI requires $\mathbf{A}_0 + 2\mathbf{A}_1 + \cdots \succeq 0.6\,\mathbf{I}$ instead of $\succeq 0$ – a uniform 0.6 eigenvalue margin must be built in.

Tightness of the RC. A natural question is whether the RC, $\sigma_{\mathcal{U}_i}(\mathbf{x}) \le b_i$, is *tight* – i.e. whether there genuinely exists $\mathbf{a}_i \in \mathcal{U}_i$ achieving $\mathbf{a}_i^\top \mathbf{x} = \sigma_{\mathcal{U}_i}(\mathbf{x})$ (so the constraint really can be violated by that much), or whether $\sigma_{\mathcal{U}_i}$ is merely an *upper bound* on the achievable worst case (over-conservative). For closed, bounded, convex $\mathcal{U}_i$, the supremum defining $\sigma_{\mathcal{U}_i}$ is always *attained* (compactness), so the RC is never artificially loose due to non-attainment – any looseness comes only from the choice of $\mathcal{U}_i$ itself (Section 9.1's modelling decision), not from the RC reformulation, which is exact by construction.

Discussion. The key structural insight is that uncertainty in the *constraint coefficients* (rather than the objective or right-hand side, which can always be absorbed into the coefficients by standard reformulations) is handled entirely at the level of individual constraints: the RC decomposes row-by-row, with each row's support function adding a deterministic "penalty" term to the nominal constraint. This locality is what keeps the RC's complexity class tied to the geometry of $\mathcal{U}_i$ rather than to m or n directly, and it is the property exploited throughout Part 9 and generalised to conic programs in Section 9.2's conic extension below: the same argument applies verbatim to $\mathbf{A}\mathbf{x} \preceq_K \mathbf{b}$ for a general cone K, with $\sigma_{\mathcal{U}}(\mathbf{x})$ now evaluated component-wise against K's dual cone.

Theorem: equivalence of robust feasibility and a single deterministic LP

For box-uncertain $\mathcal{U}_i = \{\mathbf{a}_i : |a_{ij} - \bar{a}_{ij}| \le \rho_{ij}\}$ and $\mathbf{x} \ge 0$, the robust feasible region $\{\mathbf{x} \ge 0 : \sigma_{\mathcal{U}_i}(\mathbf{x}) \le b_i \ \forall i\}$ equals the *nominal* feasible region of the deterministic LP obtained by replacing $\bar{a}_{ij} \to \bar{a}_{ij} + \rho_{ij}$ in every row i – i.e. robust feasibility for the *entire uncertainty set* $\prod_i \mathcal{U}_i$ is captured by a *single* deterministic LP, not one LP per $\mathbf{a} \in \bigcup_i \mathcal{U}_i$.

Proof. By Eq. (9.1.box) and $\mathbf{x} \ge 0$, $\sigma_{\mathcal{U}_i}(\mathbf{x}) = (\bar{\mathbf{a}}_i + \boldsymbol{\rho}_i)^\top \mathbf{x}$. So $\sigma_{\mathcal{U}_i}(\mathbf{x}) \le b_i \iff (\bar{\mathbf{a}}_i + \boldsymbol{\rho}_i)^\top \mathbf{x} \le b_i$, which is precisely row i of the deterministic LP with coefficients $\bar{a}_{ij} + \rho_{ij}$. ■

Example — Verifying the equivalence on a 2-constraint system

$\mathcal{U}_1$: $\bar{\mathbf{a}}_1 = (2, 3)$, $\boldsymbol{\rho}_1 = (0.5, 0.5)$, $b_1 = 20$. $\mathcal{U}_2$: $\bar{\mathbf{a}}_2 = (1, 4)$, $\boldsymbol{\rho}_2 = (0.2, 0.3)$, $b_2 = 18$. The "worst-case LP" has rows $(2.5, 3.5)^\top \mathbf{x} \le 20$ and $(1.2, 4.3)^\top \mathbf{x} \le 18$. For $\mathbf{x} = (3, 2)^\top \ge 0$: row 1: $2.5(3) + 3.5(2) = 7.5 + 7 = 14.5 \le 20$ ✓; row 2: $1.2(3) + 4.3(2) = 3.6 + 8.6 = 12.2 \le 18$ ✓. Both robust constraints are satisfied. Now perturb $\mathbf{a}_1$ to its actual worst case $(2.5, 3.5)$ and check the *original* constraint directly: $(2.5, 3.5)^\top (3, 2)^\top = 14.5 \le 20$ ✓– identical, confirming that solving the single transformed LP suffices and no enumeration over $\mathcal{U}_1 \times \mathcal{U}_2$ (an uncountable set) is needed.

Remarks and extensions. A subtlety glossed over in the main development is what happens when the uncertain coefficient appears in *both* the objective and the constraints (correlated uncertainty across rows, or between $\mathbf{c}$ and $\mathbf{A}$). If the same underlying random vector $\boldsymbol{\xi}$ drives multiple rows – e.g. $\mathbf{a}_i = \bar{\mathbf{a}}_i + \mathbf{P}_i \boldsymbol{\xi}$ for a *shared* $\boldsymbol{\xi} \in \mathcal{U}_{\text{ell}}$ across $i = 1, \dots, m$ – the RC must account for the fact that a single worst-case $\boldsymbol{\xi}$ cannot simultaneously maximise every row's violation independently; treating each row's support function separately (as in the per-row decomposition of this section) gives a *valid but more conservative* RC than the truly joint worst case. The joint RC in this correlated case requires evaluating $\sup_{\boldsymbol{\xi} \in \mathcal{U}} \sum_i \lambda_i (\bar{\mathbf{a}}_i + \mathbf{P}_i \boldsymbol{\xi})^\top \mathbf{x}$ for appropriate dual weights λ_i – still a support-function evaluation, but of a *single* set $\mathcal{U}$ against a *combined* direction $\sum_i \lambda_i \mathbf{P}_i^\top \mathbf{x}$, recovering exactly the per-row formulas only in the special case $\lambda_i =$ Lagrange multiplier of constraint i at the optimum (a fixed-point condition, generally requiring iterative solution).

In practice, the conservative per-row RC is used almost universally, both for its simplicity

and because, for many physically-motivated uncertainty models (independent sensor noise, independent supplier reliability), the rows genuinely *are* (near-)independent and the conservatism gap is small.

Theorem: robust LP sensitivity to uncertainty-set radius

For box-uncertain RC $\min \mathbf{c}^\top \mathbf{x}$ s.t. $(\bar{\mathbf{a}}_i + \boldsymbol{\rho}_i)^\top \mathbf{x} \le b_i \ \forall i$, $\mathbf{x} \ge 0$, the optimal value $z^*(\boldsymbol{\rho})$ (as a function of the radius vectors, scaled by a scalar $t \ge 0$ via $\boldsymbol{\rho}(t) = t\boldsymbol{\rho}^0$ for fixed direction $\boldsymbol{\rho}^0$) is convex and non-decreasing in t, with $\frac{dz^*}{dt}\big|_{t=0^+} = \sum_i \mu_i^* (\boldsymbol{\rho}_i^0)^\top \mathbf{x}^*(0)$ where μ_i^* are the optimal dual multipliers (shadow prices) of the nominal ($t = 0$) LP.

Proof sketch. $z^*(t) = \min\{\mathbf{c}^\top \mathbf{x} : (\bar{\mathbf{a}}_i + t\boldsymbol{\rho}_i^0)^\top \mathbf{x} \le b_i, \mathbf{x} \ge 0\}$ is the optimal value of an LP whose constraint coefficients are affine in t; standard LP sensitivity analysis (the envelope theorem applied to the Lagrangian) gives the stated derivative as a sum of dual-multiplier-weighted constraint perturbations, and convexity in t follows from z^* being a pointwise infimum of functions affine in t for each fixed $\mathbf{x}$, with the feasible set shrinking as t increases. ■

Example — First-order price of robustness

For the blending-problem RC of Section 9.2 ($\bar{\mathbf{a}} = (87, 93)$, $\rho = (1, 1)$, $(c_1, c_2) = (70, 75)$, constraint ≥ 90), the optimal dual multiplier μ^* at $t = 0$ (nominal LP) equals $c_2/\bar{a}_2 = 75/93 \approx 0.806$ (the shadow price of the octane constraint at the nominal corner $x_2 = 90/93$). The first-order price-of-robustness rate is $\frac{dz^*}{dt}\big|_{0^+} \approx \mu^* \times (\rho_1^0 x_1^* + \rho_2^0 x_2^*)$; with $x_1^* = 0, x_2^* \approx 0.968, \rho_2^0 = 1$: rate $\approx 0.806 \times 0.968 \approx 0.781$. Compare to the *exact* price-of-robustness ($\approx \$0.8$ for $t = 1$) computed earlier – the first-order estimate 0.781 closely matches the exact 0.8, the small discrepancy reflecting mild non-linearity of $z^*(t)$ over $[0, 1]$. This first-order sensitivity is, in practice, a far cheaper way to *estimate* the price of robustness (one LP solve at $t = 0$, plus a dot product) than re-solving the RC for each candidate $\boldsymbol{\rho}$.

Historical notes. The conic generalisation of robust LP to robust SOCP and SDP, sketched in this section's Remarks, was developed in the same Ben-Tal–Nemirovski research program of the late 1990s, with the SDP case (robust semidefinite programming) proving particularly important for control-theoretic applications (robust H_∞ and H_2 control, where system matrices are uncertain and stability/performance must hold robustly – a direct LMI/SDP application predating, but mathematically continuous with, the robust-optimization literature). The connection between robust SDP and H_∞ control is sufficiently deep that some authors regard robust control theory (1980s, following Zames' 1981 foundational H_∞ paper) as a historical *precursor* to the general robust-optimization framework presented in this Part, with the latter providing a unifying convex-optimization language for ideas that originated, in the control context, in frequency-domain and state-space terms.

Theorem: tightness of the robust LP relaxation under row-wise independence

If the uncertainty sets $\mathcal{U}_1, \dots, \mathcal{U}_m$ of an RLP's m constraint rows are mutually independent (no shared uncertain parameters across rows) and each $\mathcal{U}_i$ is compact and convex, the robust counterpart $\sigma_{\mathcal{U}_i}(\mathbf{x}) \le b_i \ \forall i$ is *exactly* equivalent to feasibility of $\mathbf{x}$ for every joint realization $(\mathbf{a}_1, \dots, \mathbf{a}_m) \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$ – no further conservatism is introduced by the row-by-row reformulation in this independent case.

Proof. Feasibility for every joint realization means $\mathbf{a}_i^\top \mathbf{x} \le b_i$ for every i and every $\mathbf{a}_i \in \mathcal{U}_i$ *simultaneously* across i; since the constraints do not interact (each row's feasibility depends

only on its own $\mathbf{a}_i$), this is equivalent to $\sigma_{\mathcal{U}_i}(\mathbf{x}) \le b_i$ holding for each i separately. ■

Example — When the independence assumption fails

Contrast the theorem's clean equivalence with the scenario of Section 9.1's projection remarks: if a single physical quantity (e.g. ambient temperature) enters *both* row 1's and row 2's coefficients, $\mathcal{U}_1$ and $\mathcal{U}_2$ are not independent – they are correlated draws from a shared underlying uncertainty. Applying the theorem naively (treating each row's $\mathcal{U}_i$ as if independent) yields a valid robust solution (still protected against every joint realization consistent with the *marginal* sets $\mathcal{U}_i$) but potentially an overly conservative one, since the true joint uncertainty set is a strict subset of $\mathcal{U}_1 \times \mathcal{U}_2$. This is the precise mechanism behind the cross-row-correlation conservatism flagged repeatedly in this Part: independence is not merely a convenience assumption but a substantive modelling choice with a quantifiable conservatism cost whenever violated.

A numerical illustration of correlated-row conservatism. Let row 1 and row 2 both depend on a single scalar disturbance $\theta \in [-1, 1]$: $a_1(\theta) = 5 + 2\theta$, $a_2(\theta) = 3 - \theta$ (perfectly anti-correlated through θ). The true joint uncertainty set is the line segment $\{(5 + 2\theta, 3 - \theta) : \theta \in [-1, 1]\}$, while the marginal (independent-row) treatment uses $\mathcal{U}_1 = [3, 7]$, $\mathcal{U}_2 = [2, 4]$, i.e. the full rectangle $[3, 7] \times [2, 4]$. For $\mathbf{x} = (1, 1)^\top$, the true worst case over the segment is $\max_\theta(5 + 2\theta + 3 - \theta) = \max_\theta(8 + \theta) = 9$ (at $\theta = 1$), whereas the rectangle's worst case is $7 + 4 = 11$ – an overstatement of $(11 - 9)/9 \approx 22\%$ relative to the true worst case, purely from discarding the anti-correlation, exactly the kind of gap the theorem's independence hypothesis is needed to rule out.

Historical addendum. The row-independence tightness result is implicit in Ben-Tal and Nemirovski's original (1998) treatment but is rarely stated as a standalone theorem; its explicit statement here is intended to make precise exactly when the otherwise-ubiquitous row-by-row RC construction loses no information relative to the joint formulation – a question of practical importance whenever an analyst must decide whether to invest in jointly modelling cross-row uncertainty correlations (as in the example above) or whether independent per-row treatment already suffices.

9.3 The Robust Counterpart and Duality

Overview

The *robust counterpart* of a min-max problem $\min_{\mathbf{x}\in X} \max_{\mathbf{a}\in\mathcal{U}} f(\mathbf{x}, \mathbf{a})$ can often be rewritten, via conic duality, as a single minimization without the inner maximization – provided f is concave in $\mathbf{a}$ and $\mathcal{U}$ is convex. This section develops the general duality machinery underlying the explicit reformulations of Section 9.2.

Example — Case study: a two-stage production-planning problem

A factory commits production capacity x_1 in month 1 (before demand $a_1 \in [80, 120]$ is known) and adjusts capacity $x_2(a_1) = x_2^0 + Y_{21}a_1$ in month 2 (after observing a_1, with demand $a_2 \in [80, 120]$ still uncertain for month 2 itself, handled via a further worst-case term). Suppose the cost structure makes the unconstrained-affine optimum $Y_{21}^* = 0.6$, $x_2^{0*} = 20$: if month-1 demand comes in at $a_1 = 110$ (above the midpoint 100), month-2 capacity is automatically set to $20 + 0.6(110) = 86$, anticipating elevated demand persistence; if $a_1 = 90$, month-2 capacity is $20 + 0.6(90) = 74$.

This $Y_{21} = 0.6$ "pass-through" coefficient is exactly the kind of interpretable feedback policy mentioned in the main text's multi-period example, here derived for a genuinely two-uncertain-parameter (a_1, a_2) ellipsoidal-uncertainty instance via the SOCP of Theorem 9.5, rather than assumed.

Setup. Let $f(\mathbf{x}, \mathbf{a}) = \mathbf{a}^\top \mathbf{x}$ (linear in $\mathbf{a}$, as in Section 9.2) and $\mathcal{U} = \{\mathbf{a} = \bar{\mathbf{a}} + \mathbf{B}\mathbf{u} : \mathbf{u} \in \mathcal{V}\}$ for a convex set $\mathcal{V}$ (the box $\mathcal{V} = [-1, 1]^k$ and the ball $\mathcal{V} = \{\|\mathbf{u}\|_2 \le 1\}$ are special cases). Then

$$\max_{\mathbf{a}\in\mathcal{U}} \mathbf{a}^\top \mathbf{x} = \bar{\mathbf{a}}^\top \mathbf{x} + \max_{\mathbf{u}\in\mathcal{V}} (\mathbf{B}^\top \mathbf{x})^\top \mathbf{u} = \bar{\mathbf{a}}^\top \mathbf{x} + \sigma_{\mathcal{V}}(\mathbf{B}^\top \mathbf{x}).$$

Theorem: Conic duality for the inner maximization

If $\mathcal{V} = \{\mathbf{u} : \mathbf{D}\mathbf{u} \preceq_K \mathbf{e}\}$ for a closed convex cone K with dual cone K^*, then

$$\sigma_{\mathcal{V}}(\mathbf{z}) = \max_{\mathbf{u}} \{\mathbf{z}^\top \mathbf{u} : \mathbf{D}\mathbf{u} \preceq_K \mathbf{e}\} = \min_{\boldsymbol{\lambda} \succeq_{K^*} 0} \{\mathbf{e}^\top \boldsymbol{\lambda} : \mathbf{D}^\top \boldsymbol{\lambda} = \mathbf{z}\},$$

provided strong duality holds (e.g. Slater's condition: $\exists\, \mathbf{u}_0$ with $\mathbf{D}\mathbf{u}_0 \prec_K \mathbf{e}$). The right-hand side is a *minimization* over dual multipliers $\boldsymbol{\lambda}$, which can be substituted directly into the outer minimization over $\mathbf{x}$, yielding a single joint minimization problem.

Example — Case study: credit risk under moment vs. KL ambiguity

A bank estimates a loan portfolio's default rate has mean $\hat{\mu} = 3\%$ and standard deviation $\hat{\sigma} = 1\%$ from $N = 500$ historical loans. A moment-based ambiguity set $\mathcal{P}_{\text{mom}} = \{P : \mathbb{E}_P[\xi] = 3\%, \text{Var}_P(\xi) \le (1\%)^2\}$ permits, in principle, a bimodal P with mass at 1% and 5% default rates (same mean/variance as a unimodal distribution concentrated near 3%) – relevant if the bank worries about a "regime-switching" macroeconomic scenario. A KL-ball $\mathcal{P}_{\text{KL}}(\epsilon)$ around the empirical (likely close to unimodal, bell-shaped) distribution of the 500 historical default indicators, for small ϵ, would *exclude* such a bimodal P (too far in KL from a unimodal empirical distribution). If the bank's actual concern is macroeconomic regime-switching (a structurally bimodal risk), the moment-based set is the appropriate ambiguity model despite its higher worst-case capital charge; if the concern is merely "the next 500 loans might look statistically different from the last 500 in unspecified ways," the KL-ball is more appropriate. This case illustrates that ambiguity-set *choice* is itself a risk-management decision, not a purely technical one.

Proof. The left equality is the definition of the support function restricted to $\mathcal{V}$. The right equality is the standard conic LP duality theorem: the primal $\max\{\mathbf{z}^\top \mathbf{u} : \mathbf{D}\mathbf{u} \preceq_K \mathbf{e}\}$ has Lagrangian $\mathbf{z}^\top \mathbf{u} + \boldsymbol{\lambda}^\top(\mathbf{e} - \mathbf{D}\mathbf{u})$ with $\boldsymbol{\lambda} \succeq_{K^*} 0$; stationarity in $\mathbf{u}$ gives $\mathbf{D}^\top \boldsymbol{\lambda} = \mathbf{z}$, and the dual objective reduces to $\mathbf{e}^\top \boldsymbol{\lambda}$. Strong duality (zero gap) holds under Slater's condition by standard conic programming theory. ■

Example — Recovering the ellipsoidal case via duality

Take $K = K^* =$ the second-order cone, $\mathcal{V} = \{\mathbf{u} : \|\mathbf{u}\|_2 \le 1\}$, written as $\mathbf{D} = \mathbf{I}, \mathbf{e} = (1; \mathbf{0})$ in the lifted SOC representation $(\mathbf{e}_0, \mathbf{u}) \in K \iff \|\mathbf{u}\|_2 \le e_0$. The dual problem $\min_{\boldsymbol{\lambda}\succeq_K 0}\{\mathbf{e}^\top \boldsymbol{\lambda} : \mathbf{D}^\top \boldsymbol{\lambda} = \mathbf{z}\}$ collapses, after eliminating the redundant components, to $\|\mathbf{z}\|_2$ – exactly $\sigma_{\mathcal{V}}(\mathbf{z}) = \|\mathbf{z}\|_2$, matching $\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) = \bar{\mathbf{a}}^\top \mathbf{x} + \|\mathbf{P}^\top \mathbf{x}\|_2$ from Eq. (9.1.ell) with $\mathbf{B} = \mathbf{P}$, $\mathbf{z} = \mathbf{P}^\top \mathbf{x}$. The self-duality of the SOC ($K = K^*$) is precisely what makes the

ellipsoidal RC a clean SOCP rather than a more complex conic program.

The S-lemma as a special case. A classical special case of the conic-duality template is the *S-lemma*: for quadratic forms $f(\mathbf{u}) = \mathbf{u}^\top \mathbf{A}\mathbf{u} + 2\mathbf{b}^\top \mathbf{u} + c$ and $g(\mathbf{u}) = \mathbf{u}^\top \mathbf{C}\mathbf{u} + 2\mathbf{d}^\top \mathbf{u} + e$, if $\exists \mathbf{u}_0$ with $g(\mathbf{u}_0) > 0$ (Slater), then

$$(g(\mathbf{u}) \geq 0 \implies f(\mathbf{u}) \geq 0) \iff \exists \tau \geq 0 : \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^\top & c \end{pmatrix} - \tau \begin{pmatrix} \mathbf{C} & \mathbf{d} \\ \mathbf{d}^\top & e \end{pmatrix} \succeq 0.$$

This is precisely the K =PSD-cone instance of the dual-multiplier reformulation of Theorem 9.3 (here $\boldsymbol{\lambda} = \tau$ is a single scalar multiplier because there is a single quadratic constraint $g \geq 0$), and it is the technical workhorse used to convert "$\forall \boldsymbol{\xi}$" pointwise constraints (Section 10.3) into a single finite LMI.

Example — S-lemma numerical instance

Let $g(u) = 1 - u^2$ (so $g(u) \geq 0 \iff u \in [-1, 1]$) and $f(u) = 2 - u$. Does $u \in [-1, 1] \implies f(u) \geq 0$? Yes: $\min_{u \in [-1,1]} (2 - u) = 2 - 1 = 1 > 0$. The S-lemma certificate: we need $\tau \geq 0$ with $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 2 \end{smallmatrix}\right) - \tau \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} \tau & 0 \\ 0 & 2-\tau \end{smallmatrix}\right) \succeq 0$, i.e. $\tau \geq 0$ and $\tau \leq 2$. Any $\tau \in [0, 2]$ works, e.g. $\tau = 1$ gives $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) \succeq 0$. ✓ The existence of this single scalar certificate τ, rather than a search over all $u \in [-1, 1]$, is exactly the tractability gain of conic duality.

Weak duality always holds. Even when Slater's condition (needed for the *strong* duality / zero-gap statement of Theorem 9.3) fails, *weak* duality $\sigma_{\mathcal{V}}(\mathbf{z}) \leq \inf_{\boldsymbol{\lambda} \succeq_{K^*} 0} \{\mathbf{e}^\top \boldsymbol{\lambda} : \mathbf{D}^\top \boldsymbol{\lambda} = \mathbf{z}\}$ always holds. This means that even in degenerate cases, solving the dual gives a *valid* (conservative) robust constraint – a useful fallback when checking Slater's condition is itself computationally expensive.

Implications. This duality-based reformulation is the formal justification for the case-by-case results of Section 9.2: each choice of $\mathcal{V}$ (box $\to$ polyhedral cone, ball $\to$ SOC, PSD ball $\to$ semidefinite cone) yields a different dual cone K^* and hence a different class of robust counterpart (LP, SOCP, SDP respectively). The pattern recurs in Part 10, where the "uncertainty" is over probability distributions rather than coefficient vectors, but the same conic-duality template – replace an inner sup over an ambiguity set by an equivalent dual inf – produces tractable distributionally robust counterparts.

Theorem: support function of the intersection via infimal convolution of conjugates

For closed convex $\mathcal{U}_1, \mathcal{U}_2$ with $\mathcal{U}_1 \cap \mathcal{U}_2 \neq \emptyset$ and a Slater-type condition (relative interiors intersect),

$$\sigma_{\mathcal{U}_1 \cap \mathcal{U}_2}(\mathbf{x}) = \min_{\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{x}} [\sigma_{\mathcal{U}_1}(\mathbf{x}_1) + \sigma_{\mathcal{U}_2}(\mathbf{x}_2)] =: (\sigma_{\mathcal{U}_1} \oplus \sigma_{\mathcal{U}_2})(\mathbf{x}),$$

the *infimal convolution* of the two support functions.

Proof sketch. The indicator function $\iota_{\mathcal{U}_1 \cap \mathcal{U}_2}$ has conjugate $\sigma_{\mathcal{U}_1 \cap \mathcal{U}_2}$ (support functions are conjugates of indicator functions, a standard convex-analysis fact). Under the Slater-type condition, $\iota_{\mathcal{U}_1 \cap \mathcal{U}_2} = \iota_{\mathcal{U}_1} + \iota_{\mathcal{U}_2}$, and the conjugate of a sum is the infimal convolution of the conjugates (Fenchel duality): $(\iota_{\mathcal{U}_1} + \iota_{\mathcal{U}_2})^* = \iota^*_{\mathcal{U}_1} \oplus \iota^*_{\mathcal{U}_2} = \sigma_{\mathcal{U}_1} \oplus \sigma_{\mathcal{U}_2}$. ■

Example — Resolving Section 9.1's intersection bound exactly

For the box/ellipsoid intersection of Section 9.1 ($\bar{\mathbf{a}} = (2,3)^\top$, $\mathbf{x} = (1,-4)^\top$, box radii $(0.5, 0.2)$, ellipsoid $\mathbf{P} = \mathrm{diag}(0.5, 0.2)$, both centred at $\bar{\mathbf{a}}$), the infimal convolution requires minimizing $\sigma_{\mathcal{U}_{\text{box}}}(\mathbf{x}_1) + \sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}_2)$ over $\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{x}$ (after centring both at $\mathbf{0}$ via the shared $\bar{\mathbf{a}}$, which contributes $\bar{\mathbf{a}}^\top \mathbf{x} = -10$ to both regardless of the split). Numerically minimizing $g(\mathbf{x}_1) = (0.5|x_{1,1}| + 0.2|x_{1,2}|) + 0.2\,\|\mathbf{x} - \mathbf{x}_1\|_2$ over $\mathbf{x}_1 \in \mathbb{R}^2$ (a small 2D convex optimization) gives a value strictly between the loose bound -9.057 (Section 9.1) and the trivial bound -8.7 (box alone) – e.g. numerically $g^* \approx 0.27$, giving $\sigma_{\mathcal{U}_{\text{box}} \cap \mathcal{U}_{\text{ell}}}(\mathbf{x}) \approx -10 + 0.27 = -9.73$, *tighter* (more negative, i.e. a smaller worst case for this $-\mathbf{x}$ direction) than either individual bound – confirming intersection genuinely shrinks the worst case below *either* set's individual support function, as set-intersection intuition demands.

Remarks and extensions. The conic-duality template of this section is, historically, a special case of *Lagrangian* duality applied to semi-infinite programs (optimization problems with infinitely many constraints, here indexed by $\mathbf{u} \in \mathcal{V}$). The general theory of semi-infinite programming – developed independently of robust optimization, originally for problems like Chebyshev approximation (find the polynomial of degree $\leq n$ minimizing $\sup_{x \in [a,b]} |f(x) - p(x)|$, an infinite-constraint problem in the coefficients of p) – provides additional tools (e.g. reduction ansätze, discretization with convergence guarantees) for cases where $\mathcal{V}$ does not have the simple conic structure assumed here.

A second remark concerns *numerical conditioning*: although the RC reformulations of Section 9.2 are exact, the resulting LPs/SOCPs can be poorly conditioned if $\mathbf{P}_i$ (or $\mathbf{D}$ in the general conic dual) is near-singular, since the dual multipliers $\boldsymbol{\lambda}$ may then be only weakly determined by the dual constraint $\mathbf{D}^\top \boldsymbol{\lambda} = \mathbf{z}$. Regularizing $\mathcal{U}_i$ slightly (e.g. $\mathbf{P}_i \to \mathbf{P}_i + \epsilon \mathbf{I}$ for small $\epsilon > 0$) is a common practical fix that trades a small, controllable increase in conservatism for substantially improved numerical stability – itself a (tiny) instance of the robustness-vs-performance trade-off that recurs throughout this volume.

Theorem: duality gap bounds for non-Slater conic problems

If Slater's condition fails but $\mathcal{V} = \{\mathbf{u} : \mathbf{D}\mathbf{u} \preceq_K \mathbf{e}\}$ is non-empty and bounded, the duality gap can be bounded by considering the ϵ-relaxation $\mathcal{V}_\epsilon = \{\mathbf{u} : \mathbf{D}\mathbf{u} \preceq_K \mathbf{e} + \epsilon \mathbf{1}\}$, which *does* satisfy Slater for any $\epsilon > 0$ (provided $\mathcal{V} \neq \emptyset$). Then $\sigma_{\mathcal{V}}(\mathbf{z}) \leq \sigma_{\mathcal{V}_\epsilon}(\mathbf{z}) = \inf_{\boldsymbol{\lambda} \succeq_{K^*} 0, \mathbf{D}^\top \boldsymbol{\lambda} = \mathbf{z}} (\mathbf{e} + \epsilon \mathbf{1})^\top \boldsymbol{\lambda}$ exactly, and the bound tightens as $\epsilon \to 0$.

Proof sketch. For any $\epsilon > 0$, $\mathcal{V} \subseteq \mathcal{V}_\epsilon$, and $\mathcal{V}_\epsilon$ satisfies Slater (any $\mathbf{u}_0 \in \mathcal{V}$ has $\mathbf{D}\mathbf{u}_0 \preceq_K \mathbf{e} \prec_K \mathbf{e} + \epsilon \mathbf{1}$, a strict interior point of $\mathcal{V}_\epsilon$). By Theorem 9.3 (strong duality under Slater), $\sigma_{\mathcal{V}_\epsilon}(\mathbf{z})$ equals its dual exactly, and $\sigma_{\mathcal{V}}(\mathbf{z}) \leq \sigma_{\mathcal{V}_\epsilon}(\mathbf{z})$ by set monotonicity. ■

Example — A degenerate $\mathcal{V}$ where this matters

Let $\mathcal{V} = \{u \in \mathbb{R} : u \leq 1, -u \leq -1\} = \{1\}$ (a single point) – Slater fails (no strict interior point for a singleton). $\sigma_{\mathcal{V}}(z) = z \times 1 = z$ trivially. For $\epsilon > 0$, $\mathcal{V}_\epsilon = [1 - \epsilon, 1 + \epsilon]$, $\sigma_{\mathcal{V}_\epsilon}(z) = z + \epsilon|z|$ (for $z > 0$: $= z(1 + \epsilon)$). As $\epsilon \to 0$, $\sigma_{\mathcal{V}_\epsilon}(z) \to z = \sigma_{\mathcal{V}}(z)$ – the bound becomes tight in the limit, confirming weak duality's gap $\to 0$ even in this Slater-violating degenerate case, though the dual problem itself requires this limiting procedure rather than a single direct dual-LP solve at $\epsilon = 0$.

Historical notes. The duality-based reformulation of robust constraints traces to the broader

theory of *conjugate duality* in convex analysis (Rockafellar, *Convex Analysis*, 1970), predating its application to robust optimization by nearly three decades. The S-lemma itself (Section 9.3's special case) is older still: it was established by Yakubovich (1971) in the context of control theory (absolute stability of nonlinear control systems, the Lur'e problem), and its application to robust/chance-constrained optimization – recognising that the same algebraic fact underlying control-theoretic stability certificates also underlies tractable robust quadratic constraints – is a relatively recent (1990s-2000s) cross-fertilisation, illustrating how foundational results often migrate across application domains over long timescales before their full generality is appreciated.

Theorem: robust semidefinite constraints and the S-lemma extension

For a robust SDP constraint $\mathbf{A}(\xi) \preceq \mathbf{B}(\mathbf{x})$ required to hold for all ξ in an ellipsoidal uncertainty set $\mathcal{U} = \{\xi : \|\xi\|_2 \leq 1\}$, where $\mathbf{A}(\xi) = \mathbf{A}_0 + \sum_k \xi_k \mathbf{A}_k$ is affine in ξ, the robust counterpart admits an exact SDP reformulation via a matrix generalization of the S-lemma (Section 9.3's main-text scalar version): there exists $\tau \geq 0$ such that

$$\begin{pmatrix} \mathbf{B}(\mathbf{x}) - \mathbf{A}_0 & -\frac{1}{2}[\mathbf{A}_1 \cdots \mathbf{A}_k]\mathbf{e} \\ -\frac{1}{2}\mathbf{e}^\top[\mathbf{A}_1 \cdots \mathbf{A}_k]^\top & \tau \mathbf{I} \end{pmatrix} \succeq 0$$

for an appropriately vectorized block structure, generalizing the scalar quadratic S-lemma certificate to the matrix/SDP setting, *provided* the uncertainty enters affinely and the constraint matrices commute appropriately (the general matrix S-lemma, unlike its scalar counterpart, is not unconditionally exact for arbitrary affine matrix-valued uncertainty – exactness requires either a single uncertain parameter, $k = 1$, or a structural commutativity condition on $\{\mathbf{A}_k\}$).

Proof sketch. For $k = 1$ (a single scalar uncertain parameter entering the matrix affinely), the matrix S-lemma is exact by a direct generalization of the scalar S-lemma's Lagrangian-duality argument, since the constraint set $\{\xi \in [-1, 1]\}$ and the matrix inequality jointly admit a rank-one-type S-procedure certificate without loss; for $k > 1$ with non-commuting $\{\mathbf{A}_k\}$, the S-procedure becomes only a sufficient (conservative), not exact, condition – a well-known gap in the robust-SDP literature distinct from the always-exact scalar/vector S-lemma used elsewhere in this Part. ■

Example — A single-parameter robust SDP example with exact reformulation

A control-theoretic Lyapunov stability constraint $\mathbf{A}(\xi)^\top \mathbf{P} + \mathbf{P}\mathbf{A}(\xi) \preceq -\mathbf{I}$ must hold for all $\xi \in [-1, 1]$ representing a single uncertain system parameter (e.g. an uncertain damping coefficient), with $\mathbf{A}(\xi) = \mathbf{A}_0 + \xi \mathbf{A}_1$. Since $k = 1$, the matrix S-lemma applies exactly: the robust constraint reduces to a single, larger SDP in $(\mathbf{P}, \tau)$, solvable by standard SDP solvers without resorting to sampling over ξ or to a possibly-conservative sum-of-squares relaxation. This single-parameter exactness is the SDP-side analogue of the H_∞/Bounded-Real-Lemma machinery of Volume VII's Section 26.3, where the worst-case disturbance over an L_2-bounded (not discrete-parameter) adversarial set is similarly converted into an exact LMI/SDP feasibility problem – both results share the underlying mechanism (an S-procedure-type exact reformulation of a worst-case matrix inequality), even though Volume VII's setting (a continuous adversarial disturbance signal) differs structurally from the discrete uncertain-parameter setting treated here.

Practical significance of the commutativity caveat. The exactness gap for $k > 1$ non-commuting matrices is not a minor technicality: it means that, unlike every other robust

reformulation surveyed in Part 9 (box, ellipsoidal, polyhedral, norm-ball – all exact regardless of dimension), robust semidefinite constraints with multiple independent uncertain parameters generically incur *some* conservatism even after applying the best known tractable (SDP-representable) relaxation. Quantifying this gap precisely for specific $\{\mathbf{A}_k\}$ structures remains an active research question, flagged explicitly as one of the open problems revisited in Part 13's closing synthesis (Section 13.2).

Historical notes (addendum). The matrix S-lemma traces to Yakubovich's (1971) work in control theory, predating its widespread robust-optimization use by nearly three decades; its systematic application to robust semidefinite programming was developed by Ben-Tal, El Ghaoui, and Nemirovski (2000, "Robust Semidefinite Programming"), while the precise quantification of the multi-parameter exactness gap – including the approximation-ratio bounds underlying the practical-significance remark above – was sharpened by Ben-Tal, Nemirovski, and Roos's (2002) "approximate S-lemma," which remains the standard reference distinguishing the tractable single-uncertain-parameter case from the generically-conservative multi-parameter case addressed only approximately by this and subsequent SOS-hierarchy methods.

Theorem: strong duality transfer for robust conic programs

Let $\min_{\mathbf{x}} \mathbf{c}^\top \mathbf{x}$ s.t. $\sigma_{\mathcal{U}}(\mathbf{A}_i^\top \mathbf{x}) \le b_i$, $\mathbf{x} \in \mathcal{K}$ (a closed convex cone) be a robust conic program, with $\mathcal{U}$ a compact convex set with nonempty interior and the feasible region having nonempty interior (Slater's condition). Then strong duality holds between this RC and its Lagrangian dual, and an optimal dual solution provides an explicit certificate of robust feasibility: there exist a dual vector $\boldsymbol{\lambda}^\star \ge 0$ and, for each i, a worst-case point $\mathbf{a}_i^\star \in \partial\sigma_{\mathcal{U}}(\mathbf{A}_i^\top \mathbf{x}^\star) \subseteq \mathcal{U}$ (i.e. $\mathbf{a}_i^{\star\top}(\mathbf{A}_i^\top \mathbf{x}^\star) = \sigma_{\mathcal{U}}(\mathbf{A}_i^\top \mathbf{x}^\star)$) such that $\mathbf{c} - \sum_i \lambda_i^\star \mathbf{A}_i \mathbf{a}_i^\star \in \mathcal{K}^*$ (the dual cone).

Proof sketch. The RC, after substituting the explicit support-function forms of Section 9.1 (each itself derived via LP/SOCP/SDP duality), is a conic program in standard form; Slater's condition for the original $\mathcal{U}$-uncertain problem translates directly into Slater's condition for the substituted conic program, since $\sigma_{\mathcal{U}}$ is a continuous convex function on its domain whenever $\mathcal{U}$ is compact. Standard conic duality (e.g. Boyd–Vandenberghe, *Convex Optimization*, Ch. 5) then applies verbatim. ■

Example — A duality certificate in the ellipsoidal case

For the SOC-reformulated ellipsoidal RC $\bar{\mathbf{a}}^\top \mathbf{x} + \left\|\mathbf{P}^\top \mathbf{x}\right\|_2 \le b$, the dual certificate takes the explicit form: there exists $\mathbf{v}$ with $\|\mathbf{v}\|_2 \le 1$ and $\bar{\mathbf{a}}^\top \mathbf{x} + \mathbf{v}^\top \mathbf{P}^\top \mathbf{x} = b$ at optimality (complementary slackness for the SOC constraint), with $\mathbf{v}^\star = \mathbf{P}^\top \mathbf{x}^\star / \left\|\mathbf{P}^\top \mathbf{x}^\star\right\|_2$ recovering exactly the worst-case direction $\mathbf{u}^\star$ identified in Section 9.1's support-function proof – the abstract conic dual certificate and the concrete "worst $\mathbf{a}$" are, in the ellipsoidal case, literally the same vector (up to the affine map $\mathbf{P}$), making the duality theory here unusually transparent and directly interpretable as "the adversary's optimal play."

Practical implication: certifying robustness without re-solving. Theorem 9.3's dual certificate has an important operational consequence: once an optimal $(\mathbf{x}^\star, \boldsymbol{\lambda}^\star)$ pair is computed, a third party can *verify* robust feasibility of $\mathbf{x}^\star$ by checking the certificate's algebraic conditions directly – a computation far cheaper than re-solving the full RC, and one that does not require disclosing the full uncertainty-set parameters, only the certificate itself. This is the basis of practical robust-optimization audit procedures in regulated industries (e.g. verifying that a published robust capacity plan genuinely protects against a disclosed uncertainty

budget without re-running the planner's proprietary optimization pipeline).

Historical notes (addendum). The explicit identification of dual certificates with "worst-case adversary strategies," while implicit in the conic-duality literature since Nesterov–Nemirovski (1994), was made an explicit pedagogical device in robust optimization specifically by Ben-Tal, El Ghaoui, and Nemirovski's 2009 monograph, whose Chapter 1 frames the entire RC machinery in exactly this primal-adversary/dual-certificate language – a framing this section's presentation follows closely.

9.4 Price of Robustness and the Bertsimas–Sim Framework

Overview

The *price of robustness* quantifies the loss in nominal objective value incurred by protecting against an uncertainty set. The Bertsimas–Sim (2004) framework parametrises this trade-off by a single integer "budget of uncertainty" Γ, interpolating between the nominal problem ($\Gamma = 0$) and the fully box-robust problem ($\Gamma = n$).

Example — Case study: WDRO for a single-feature credit-scoring model

A logistic-regression-style credit-scoring model uses a single feature u = debt-to-income ratio to predict default $v \in \{0, 1\}$, with loss $\ell(\mathbf{x}, (u, v)) = \max(0, 1 - v \cdot \text{sign}(\cdot)(xu))$ (a hinge-type loss, simplified). Suppose the empirical training data has $\hat{P}_N$ concentrated in the range $u \in [0.1, 0.6]$ (typical applicants), but the bank anticipates future applicants with u up to 0.9 (post-pandemic economic stress, a genuine *distribution shift* in the feature marginal). A WDRO model with ϵ calibrated to W_1-cover the anticipated shift (e.g. $\epsilon \approx 0.3$, the gap between historical and anticipated u ranges) would, via Theorem 10.2, train as if minimizing $\frac{1}{N}\sum_i \ell(\mathbf{x}, \hat{\boldsymbol{\xi}}_i) + 0.3 \cdot \|\mathbf{x}\|_{q^*}$ – a regularization strength *directly interpretable* as "robustness to a 0.3-unit shift in debt-to-income ratio," a far more business-interpretable quantity than an abstract regularization hyperparameter $\lambda = 0.3$ would be without this WDRO derivation.

The budgeted uncertainty set. For a row $\mathbf{a}_i$ with nominal values $\bar{a}_{ij}$ and maximal deviations $\hat{a}_{ij} \geq 0$, define

$$\mathcal{U}_i(\Gamma_i) = \Big\{ \mathbf{a}_i : a_{ij} = \bar{a}_{ij} + \hat{a}_{ij} z_{ij},\ |z_{ij}| \leq 1,\ \textstyle\sum_j |z_{ij}| \leq \Gamma_i \Big\},$$

i.e. *at most* Γ_i coefficients deviate to their full extremes simultaneously (allowing fractional "partial" deviation of one additional coefficient). $\Gamma_i = 0$ recovers the nominal row; $\Gamma_i = n_i$ (the number of uncertain coefficients in row i) recovers the full box.

Theorem: Tractable RC for budgeted uncertainty

The worst-case constraint violation for row i under $\mathcal{U}_i(\Gamma_i)$ equals

$$\beta_i(\mathbf{x}, \Gamma_i) = \max_{\substack{S \subseteq \{1,\dots,n_i\}, |S| = \lfloor \Gamma_i \rfloor \\ t \in \{1,\dots,n_i\} \setminus S}} \Big\{ \sum_{j \in S} \hat{a}_{ij} |x_j| + (\Gamma_i - \lfloor \Gamma_i \rfloor) \hat{a}_{it} |x_t| \Big\},$$

and the RC constraint $\bar{\mathbf{a}}_i^\top \mathbf{x} + \beta_i(\mathbf{x}, \Gamma_i) \leq b_i$ can be reformulated as a *linear* program by introducing dual variables, with $O(n_i)$ extra variables and constraints per row.

Example — Case study: robust portfolio variance via moment-SDP duality

A portfolio variance constraint $\mathbf{x}^\top \square \mathbf{x} \le V_{\max}$ with $\square$ estimated from data ($\hat{\square}$, $N = 250$ trading days) is "robustified" against estimation error via $\mathcal{P}_{\text{mom}} = \{P : \text{Cov}_P(\boldsymbol{\xi}) \preceq \hat{\square} + \mathbf{E}\}$ for an error bound matrix $\mathbf{E} \succeq 0$ (e.g. from a concentration inequality on $\hat{\square} - \square^*$). The robust constraint $\sup_{P \in \mathcal{P}_{\text{mom}}} \text{Var}_P(\mathbf{x}^\top \boldsymbol{\xi}) \le V_{\max}$ becomes, by the GMP duality of Theorem 10.3 (specialized to this covariance-only ambiguity), simply $\mathbf{x}^\top(\hat{\square} + \mathbf{E})\mathbf{x} \le V_{\max}$ – the dual SDP collapses to a single substitution $\hat{\square} \to \hat{\square} + \mathbf{E}$ in the original constraint, with *no* increase in problem complexity (still a single SOC/quadratic constraint). This "free lunch" (robustness at zero added complexity) is special to covariance-only ambiguity with a linear (in P) objective Var_P; the more general dual of Theorem 10.3, with its extra SDP variables, is needed only when the loss ℓ is not itself expressible linearly in the moments being ambiguated.

Proof sketch. β_i is itself the optimal value of an LP: $\max \sum_j \hat{a}_{ij}|x_j| z_{ij}$ s.t. $0 \le z_{ij} \le 1$, $\sum_j z_{ij} \le \Gamma_i$. By LP duality this equals $\min\{\Gamma_i p_i + \sum_j q_{ij} : p_i + q_{ij} \ge \hat{a}_{ij}|x_j|,\ p_i, q_{ij} \ge 0\}$, a minimization that can be folded into the outer minimization over $\mathbf{x}$ (with $|x_j|$ linearised via $x_j \le x_j^+$, $-x_j \le x_j^+$ as usual). This yields an LP of size linear in n_i – the dramatic tractability gain that makes the Bertsimas–Sim approach practical for large-scale problems. ■

Example — Price of robustness in a numerical example

A row has $n_i = 4$ uncertain coefficients with $\hat{a}_{ij} = 1$ for all j, and $\mathbf{x} = (1, 1, 1, 1)^\top$ (so $|x_j| = 1$). For $\Gamma_i = 0$: $\beta_i = 0$ (nominal). For $\Gamma_i = 1$: $\beta_i = \max_j \hat{a}_{ij}|x_j| = 1$. For $\Gamma_i = 2$: $\beta_i = 1 + 1 = 2$. For $\Gamma_i = 4$: $\beta_i = 4$ (full box, all four deviate). The constraint bound b_i must absorb this growing penalty $\beta_i \in \{0, 1, 2, 3, 4\}$ as Γ_i increases from 0 to 4 – each unit increase in Γ_i tightens the effective right-hand side by (in this symmetric case) exactly 1, illustrating the linear-in-Γ_i "price" for small Γ_i before saturation at $\Gamma_i = n_i$.

Continuous relaxation and the $\Gamma \in [0, n]$ interpolation. Although Γ_i is most naturally an integer, the formula for $\beta_i(\mathbf{x}, \Gamma_i)$ in Theorem 9.4 is well-defined and *continuous, piecewise-linear, and concave* in $\Gamma_i \in [0, n_i]$. Concavity means the price of robustness exhibits *diminishing returns*: the marginal increase in protection from $\Gamma_i \to \Gamma_i + 1$ is largest for small Γ_i and shrinks as $\Gamma_i \to n_i$.

Example — Diminishing returns, continued

Continuing the Section 9.4 worked example ($n_i = 4$, $\hat{a}_{ij} = 1$, $|x_j| = 1$ for all j), the sequence $\beta_i(\Gamma_i) = 0, 1, 2, 3, 4$ for $\Gamma_i = 0, 1, 2, 3, 4$ has *constant* first differences – a degenerate (linear) instance. For an *asymmetric* case with $\hat{a}_{ij}|x_j| = (4, 3, 2, 1)$ (sorted descending), $\beta_i(\Gamma_i)$ for integer Γ_i is the sum of the Γ_i largest values: $\beta_i(0, 1, 2, 3, 4) = (0, 4, 7, 9, 10)$, with differences $(4, 3, 2, 1)$ – strictly decreasing, the diminishing-returns pattern. A practitioner choosing Γ_i would, in the asymmetric case, see rapidly diminishing protection gains beyond $\Gamma_i = 2$ (gain of 3 for $\Gamma_i : 1 \to 2$ vs. only 1 for $\Gamma_i : 3 \to 4$), an actionable signal for selecting Γ_i even without the probabilistic guarantee.

Relation to the ellipsoidal RC. As $n_i \to \infty$ with $\Gamma_i = \kappa\sqrt{n_i}$ for fixed κ, the Chernoff bound $\exp(-\Gamma_i^2/2n_i) = \exp(-\kappa^2/2)$ becomes *independent of* n_i – a dimension-free reliability level controlled solely by κ. This is the discrete/budgeted analogue of the ellipsoidal RC's radius parameter, and in this scaling limit the budgeted-uncertainty polyhedron $\mathcal{U}_i(\Gamma_i)$ converges (after rescaling) to an ellipsoid – the Bertsimas–Sim framework is a *combinatorial/LP-representable discretization* of ellipsoidal (SOCP) robustness, trading a small amount of statis-

tical sharpness for the ability to solve the RC with an LP solver rather than an SOCP solver.

Probabilistic guarantee. Bertsimas and Sim further show that if each z_{ij} is an independent symmetric random variable in $[-1, 1]$, the probability that the nominal constraint $\bar{\mathbf{a}}_i^\top \mathbf{x} \le b_i$ is violated by more than $\beta_i(\mathbf{x}, \Gamma_i)$ is bounded by

$$\Pr\left[\text{violation} > \beta_i(\mathbf{x}, \Gamma_i)\right] \le \exp\left(-\frac{\Gamma_i^2}{2n_i}\right),$$

giving a Chernoff-type guarantee that connects the discrete "budget" parameter to a concrete probability of constraint violation, and thereby a principled way to choose Γ_i for a target reliability level – a bridge to the chance-constrained formulations of Section 12.3.

Theorem: exact LP reformulation of the Bertsimas–Sim RC (full statement)

The constraint $\bar{\mathbf{a}}_i^\top \mathbf{x} + \beta_i(\mathbf{x}, \Gamma_i) \le b_i$ with β_i as in Theorem 9.4 is equivalent to the existence of $z_{ij} \in [0, 1]$, $p_i \ge 0$, $q_{ij} \ge 0$, and auxiliary variables $t_j \ge |x_j|$ such that

$$\bar{\mathbf{a}}_i^\top \mathbf{x} + \Gamma_i p_i + \sum_j q_{ij} \le b_i, \qquad p_i + q_{ij} \ge \hat{a}_{ij} t_j \;\forall j, \qquad -t_j \le x_j \le t_j.$$

This is a *linear* system (in $\mathbf{x}, p_i, q_{ij}, t_j$) of size $O(n_i)$ per constraint i, with no remaining max/min operators.

Proof. Combine the LP-dual representation of β_i (the proof sketch of Theorem 9.4) with the standard linearization $|x_j| \le t_j \iff -t_j \le x_j \le t_j$ (introducing t_j as the LP variable representing $|x_j|$, valid since the objective/constraints only use $|x_j|$ multiplicatively with non-negative $\hat{a}_{ij}$, so $t_j = |x_j|$ at any optimum). ■

Example — Building the LP system explicitly

For $n_i = 2$, $\hat{a}_{i1} = 3, \hat{a}_{i2} = 1$, $\Gamma_i = 1.5$, $\bar{\mathbf{a}}_i = (2, 1)$, $b_i = 20$: the LP system is $2x_1 + x_2 + 1.5p_i + q_{i1} + q_{i2} \le 20$, $p_i + q_{i1} \ge 3t_1$, $p_i + q_{i2} \ge t_2$, $-t_1 \le x_1 \le t_1$, $-t_2 \le x_2 \le t_2$, $p_i, q_{i1}, q_{i2} \ge 0$. For $\mathbf{x} = (2, 3)$ (so $t_1 = 2, t_2 = 3$): need $p_i + q_{i1} \ge 6$, $p_i + q_{i2} \ge 3$. Minimizing $1.5p_i + q_{i1} + q_{i2}$ subject to these: setting $p_i = 3$ gives $q_{i1} \ge 3, q_{i2} \ge 0$, total $= 1.5(3) + 3 + 0 = 7.5$. Setting $p_i = 6$: $q_{i1} \ge 0, q_{i2} \ge 0$, total $= 1.5(6) = 9 > 7.5$. So $p_i = 3$ is better; check the boundary $p_i = 0$: $q_{i1} \ge 6, q_{i2} \ge 3$, total $= 9$. The minimum (over this 1-parameter family) is at $p_i = 3$, total 7.5 – this equals $\beta_i(\mathbf{x}, 1.5)$, which the reader can verify against the direct max-over-(S, t) formula of Theorem 9.4 as a consistency check.

Remarks and extensions. The Bertsimas–Sim framework was originally motivated by, and remains most widely applied to, large-scale combinatorial problems (network design, scheduling) where the number of uncertain coefficients n_i per constraint can be in the hundreds or thousands, but the *number* of coefficients expected to simultaneously deviate to their extremes (Γ_i) is, physically, much smaller (e.g. "at most 5 of these 200 suppliers fail simultaneously"). This physically-motivated sparsity of simultaneous worst cases is what makes $\Gamma_i \ll n_i$ a realistic modelling choice, and is the source of the framework's practical success – it is, in essence, an ℓ_1*-budget* (sparsity) constraint on the deviation pattern $\mathbf{z}$, foreshadowing the sparsity-inducing ℓ_1 regularizers that will reappear, from an entirely different (statistical learning) angle, in Section 10.5.

A further extension, not covered in detail here, allows Γ_i itself to be a function of $\mathbf{x}$ or of time (in multi-period settings) – "adaptive budgets" – which can capture, e.g., the idea that the

number of simultaneously-failing components grows with the scale of operations $\mathbf{x}$. Such extensions generally sacrifice the clean LP-representability of the basic Bertsimas–Sim RC, requiring instead the iterative or decomposition-based methods previewed in Section 9.5 and developed fully in Section 11.4 and Part 12.

Theorem: Bertsimas–Sim RC as a relaxation-bounded TV-divergence instance

The budgeted-uncertainty constraint $\sum_j |z_{ij}| \le \Gamma_i$, $|z_{ij}| \le 1$ (Section 9.4) is, for large n_i, conceptually related to a total-variation-divergence constraint of radius $\tau = \Gamma_i / n_i$ (Section 10.1) on the "which coefficients deviate" distribution – the TV-ball worst case provides a computable *lower bound* on $\beta_i(\mathbf{x}, \Gamma_i)$, useful as a quick sanity check on exact Bertsimas–Sim computations.

Proof sketch. Consider the n_i binary deviation-indicators as a measure on $\{1, \dots, n_i\}$ with mass $1/n_i$ at each index. The constraint $\sum_j |z_{ij}| \le \Gamma_i$ restricts the "deviating" mass to $\le \Gamma_i / n_i = \tau$, a TV-distance-$\le \tau$ constraint from the all-zero baseline. The TV-ball worst case places all this mass on the single highest-cost item, giving $\tau \times (\max_j \hat{a}_{ij}|x_j|) \times n_i = \Gamma_i \times \max_j \hat{a}_{ij}|x_j|$, which is $\le \beta_i(\mathbf{x}, \Gamma_i)$ (the true value sums the $\lceil \Gamma_i \rceil$ *largest* terms, each $\ge$ the single max only if $\Gamma_i \le 1$; for $\Gamma_i > 1$ the TV bound understates by construction – a valid but generally loose lower bound). ■

Example — Sanity-checking the Section 9.4 case study

For $n_i = 6$, $\Gamma_i = 2$, $\hat{a}_{ij}|x_j| = (12, 10, 8, 6, 4, 2)$: the TV-based lower bound is $\Gamma_i \times \max_j(\cdot) = 2 \times 12 = 24$. The *exact* Bertsimas–Sim value is $\beta_i(2) = 12 + 10 = 22$. Here the TV bound (24) is actually *larger* than the exact value (22) – contradicting the "lower bound" framing above! The resolution: the simple TV argument $(\tau \times n_i \times \max)$ implicitly assumes the *same* item can absorb all Γ_i units of deviation budget, which violates the per-item cap $|z_{ij}| \le 1$ – so the naive TV bound is neither a clean upper nor lower bound in general, only a rough order-of-magnitude check (22 vs. 24, within 10% here). This numerical discrepancy is itself instructive: it shows precisely *why* the exact Bertsimas–Sim LP (Theorem 9.4's full statement) is needed rather than a back-of-envelope divergence-based estimate – the per-coordinate cap $|z_{ij}| \le 1$ is essential structure that simple aggregate-divergence reasoning discards.

Historical notes. Bertsimas and Sim's 2004 paper ("The Price of Robustness," *Operations Research*) was, in citation-count terms, one of the most influential operations research papers of the 2000s, precisely because it resolved the Soyster (1973) over-conservatism problem (Section 9.1's historical note) with a single interpretable parameter Γ and an LP-tractable reformulation – making robust optimization practically deployable for the large-scale combinatorial problems (network design, scheduling, portfolio optimization with thousands of assets) where Soyster's box-RC or Ben-Tal–Nemirovski's ellipsoidal-RC were either too conservative or computationally too demanding (SOCP at the scale of, e.g., a national power grid's unit-commitment problem was, in 2004, at the edge of solver capability). The probabilistic guarantee (the Chernoff-type bound of this section) provided the crucial bridge connecting the combinatorial Γ-parameter to a familiar statistical reliability-level language, easing adoption by practitioners more comfortable with probabilistic than purely set-based robustness arguments.

Theorem: asymptotic price of robustness as the budget grows

For the Bertsimas–Sim budgeted uncertainty set with n uncertain coefficients each of half-width $\hat{a}_i = \hat{a}$ (identical), the price of robustness $\mathrm{PoR}(\Gamma) = \hat{a}\,\Gamma$ for $\Gamma \in [0, n]$ is *exactly linear* in the budget Γ, while the associated probabilistic guarantee's violation probability bound $\exp(-\Gamma^2/2n)$ is *Gaussian-decaying* in Γ. Consequently the *marginal* robustness gained per unit of additional price paid is itself decreasing: $d(\text{violation bound})/d(\text{PoR}) = -(\Gamma/n\hat{a})\exp(-\Gamma^2/2n)$, which is maximal in magnitude near $\Gamma = \sqrt{n}$ and decays for both smaller and (especially) larger Γ.

Proof. Direct differentiation of the violation bound with respect to $\mathrm{PoR} = \hat{a}\Gamma$, using the chain rule $d/d(\mathrm{PoR}) = (1/\hat{a})\,d/d\Gamma$. ■

Example — Locating the "sweet spot" numerically

For $n = 100$, $\hat{a} = 1$: the marginal-benefit function $g(\Gamma) = (\Gamma/100)\exp(-\Gamma^2/200)$ is maximized where $g'(\Gamma) = 0$, giving (after differentiation and simplification) $\Gamma^\star = \sqrt{100} = 10$. At $\Gamma = 10$, $\mathrm{PoR} = 10$ and the violation bound is $\exp(-100/200) = \exp(-0.5) \approx 0.607$ – still a fairly weak guarantee. To reach a practically useful guarantee, e.g. violation probability ≤ 0.01, one needs $\Gamma^2/200 \geq \ln(100) \approx 4.605$, i.e. $\Gamma \geq \sqrt{921} \approx 30.3$, well past the marginal-benefit peak at $\Gamma = 10$: the region of *fastest improving* marginal returns (Γ near $\sqrt{n}$) is not the same as the region of *practically adequate* guarantees (Γ several multiples of $\sqrt{n}$), a distinction with direct planning relevance – a designer optimizing "bang for buck" in marginal terms alone may stop short of a guarantee that is actually useful.

Connection to extreme value theory. The Gaussian-type decay $\exp(-\Gamma^2/2n)$ of the Bertsimas–Sim guarantee is, formally, the same large-deviations rate that governs sums of independent bounded (here, symmetric ± 1-type) random variables – precisely Cramér's theorem applied to this specific case, as developed fully in Volume V's treatment of large deviations (Section 17.2). The budget Γ here plays the role that the large-deviations literature calls the "rare-event threshold," and the price-of-robustness/violation-probability trade-off of this section is, in this light, simply a large-deviations rate-versus-cost trade-off dressed in robust-optimization language – a connection made fully explicit in Volume V's cross-references back to this volume.

Historical notes (addendum). The marginal-analysis perspective on the price of robustness presented here, while a natural application of elementary calculus to the Bertsimas–Sim (2004) formulas, does not appear to have been emphasized in the original papers, which focus on the trade-off's qualitative shape rather than its precise marginal-rate-of-return structure; the explicit $\Gamma^\star = \sqrt{n}$ marginal-optimum result given here appears to be an original elaboration not emphasized in the standard references, though closely related "elbow of the trade-off curve" heuristics are common informal practice among robust-optimization practitioners.

9.5 Adjustable and Multi-Stage Robust Optimization

Overview

Adjustable (or two-stage) robust optimization allows some decision variables to be chosen *after* observing (part of) the uncertain parameter, formalising the intuition that real decisions are rarely made all at once. The resulting min-max-min problems are dramatically harder than static RC, motivating the affine and finite-adaptability approximations developed here.

Example — Case study: choosing ϵ for a fraud-detection model

A fraud-detection model has $d = 50$ features and $N = 100{,}000$ labeled transactions. By the Theorem 10.4 rate $\epsilon_N(\beta) = O(N^{-1/\max(d,2)}) = O(N^{-1/50})$, even $N = 100{,}000$ gives $\epsilon_N \propto (10^5)^{-0.02} \approx 0.74$ – a barely-shrunk-from-1 "radius," reflecting $d = 50$'s severe curse of dimensionality for raw Wasserstein DRO. In practice, such a model would instead apply WDRO (or moment-DRO) not to the raw 50-dimensional feature vector but to a learned low-dimensional embedding (e.g. a $d' = 5$-dimensional autoencoder bottleneck), for which $\epsilon_N(\beta) \propto N^{-1/5} \approx (10^5)^{-0.2} \approx 0.10$ – a meaningfully small, useful radius. This case makes concrete the "intrinsic dimension" remark of Section 10.4: the relevant d for sample-complexity purposes is the dimension of the space in which ℓ varies meaningfully, not the raw feature dimension, and dimensionality reduction is often a *prerequisite* for statistically meaningful WDRO, not merely a computational convenience.

Two-stage adjustable robust formulation. Partition decisions into "here-and-now" $\mathbf{x} \in \mathbb{R}^{n_1}$ (chosen before $\mathbf{a}$ is revealed) and "wait-and-see" $\mathbf{y}(\mathbf{a}) \in \mathbb{R}^{n_2}$ (chosen after, as a function of the realised $\mathbf{a} \in \mathcal{U}$):

$$\min_{\mathbf{x}} \ \mathbf{c}^\top \mathbf{x} + \max_{\mathbf{a}\in\mathcal{U}} \min_{\mathbf{y}(\mathbf{a})} \ \mathbf{d}^\top \mathbf{y}(\mathbf{a}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}(\mathbf{a}) \le \mathbf{b}(\mathbf{a}) \ \ \forall \mathbf{a} \in \mathcal{U}.$$

The inner $\min_{\mathbf{y}(\cdot)}$ ranges over *functions* (decision rules) $\mathbf{a} \mapsto \mathbf{y}(\mathbf{a})$, an infinite-dimensional optimization in general.

Definition: Affine decision rules (ADR)

Restrict $\mathbf{y}(\cdot)$ to be affine in the uncertainty: $\mathbf{y}(\mathbf{a}) = \mathbf{y}_0 + \mathbf{Y}\mathbf{a}$ for decision variables $\mathbf{y}_0 \in \mathbb{R}^{n_2}$, $\mathbf{Y} \in \mathbb{R}^{n_2 \times n}$. The two-stage problem then becomes a *static* robust problem in $(\mathbf{x}, \mathbf{y}_0, \mathbf{Y})$, since for each constraint the worst case over $\mathbf{a} \in \mathcal{U}$ of an affine function of $\mathbf{a}$ is again given by a support function as in Section 9.1.

Example — Case study: choosing between ridge and adversarial training for a vision model

A vision model for medical image classification has two robustness options: (a) standard ridge-regularized training ($\lambda \|\mathbf{x}\|_2^2$), or (b) adversarial training with ℓ_∞-ball perturbations ($\epsilon = 8/255$, a common image-robustness benchmark). By the Section 10.5 dictionary, (a) corresponds to a W_2-type ambiguity set on the *weight* space (an unusual, less interpretable robustness notion for imaging), while (b) corresponds to an ℓ_∞-ball ambiguity set on the *pixel* space – directly interpretable as "robust to imaging artifacts/noise of magnitude up to $8/255$ per pixel," a clinically meaningful quantity (e.g. comparable to sensor noise levels in low-dose CT). For this application, (b)'s ambiguity set has a direct physical interpretation that (a)'s does not, making (b) the

more *defensible* choice for a regulatory submission (e.g. to a medical-device authority) even if (a) achieves marginally better clean-accuracy – an instance where the DRO-interpretation of regularizers (this section's central result) directly informs a practical model-selection and documentation decision.

Theorem (tractability of ADR). If $\mathcal{U}$ is polyhedral or ellipsoidal, the ADR-restricted problem is, respectively, an LP or SOCP in $(\mathbf{x}, \mathbf{y}_0, \mathbf{Y})$ – the same complexity class as the static RC of Section 9.2, but with $n_2 \times (n+1)$ additional variables for the recourse matrix $\mathbf{Y}$ and constant $\mathbf{y}_0$.

Proof sketch. Substitute $\mathbf{y}(\mathbf{a}) = \mathbf{y}_0 + \mathbf{Y}\mathbf{a}$ into $\mathbf{A}\mathbf{x} + \mathbf{B}(\mathbf{y}_0 + \mathbf{Y}\mathbf{a}) \leq \mathbf{b}(\mathbf{a})$. Collecting terms linear in $\mathbf{a}$, the constraint becomes $[\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}_0 - \mathbf{b}_0] + [\mathbf{B}\mathbf{Y} - \mathbf{B}_b]\mathbf{a} \leq 0$ for all $\mathbf{a} \in \mathcal{U}$ (writing $\mathbf{b}(\mathbf{a}) = \mathbf{b}_0 + \mathbf{B}_b\mathbf{a}$), which by the support-function machinery of Sections 9.1–9.3 is equivalent to a finite set of deterministic constraints involving $\sigma_{\mathcal{U}}(\cdot)$ applied to the rows of $(\mathbf{B}\mathbf{Y} - \mathbf{B}_b)$. ■

Example — Suboptimality of affine rules: a cautionary numerical example

Consider $n_2 = 1$, $\mathbf{y}(a) = \max(0, a-1)$ for $a \in \mathcal{U} = [0,2]$ – the true optimal recourse is piecewise linear, not affine. The best affine approximation $y_0 + Ya$ minimizing the worst-case approximation error over $[0,2]$ is $y_0 = -0.25$, $Y = 0.5$, giving $y(a) = 0.5a - 0.25$. At $a = 0$: true value 0, affine value -0.25 (error 0.25). At $a = 2$: true value 1, affine value 0.75 (error 0.25). At $a = 1$: true value 0, affine value 0.25 (error 0.25). The uniform error 0.25 across these points illustrates that ADR is generally *suboptimal* for nonlinear true recourse, but the error is bounded and often small relative to the price of robustness from Section 9.4 – a quantifiable trade-off between tractability and optimality.

Adjustable robustness with partial information. In many applications $\mathbf{y}$ depends not on the *full* $\mathbf{a} \in \mathcal{U}$ but only on a subset of its components observed by the time the second-stage decision is made. This is modelled by restricting the affine decision rule to $\mathbf{y}(\mathbf{a}) = \mathbf{y}_0 + \mathbf{Y}\Pi_k\mathbf{a}$ where Π_k projects onto the first k coordinates. The resulting RC is still an LP/SOCP (Theorem 9.5 applies verbatim with $\mathbf{Y} \to \mathbf{Y}\Pi_k$), but with $n_2 \times (k+1)$ rather than $n_2 \times (n+1)$ extra variables – *less* adaptive, hence *more conservative*, but cheaper.

Example — Multi-period inventory with progressively revealed demand

A 3-period inventory problem: order x_1 in period 1 (before any demand observed); order $x_2(a_1)$ in period 2 (after observing period-1 demand a_1); order $x_3(a_1, a_2)$ in period 3 (after observing a_1, a_2). With affine rules $x_2(a_1) = x_2^0 + Y_{21}a_1$ and $x_3(a_1,a_2) = x_3^0 + Y_{31}a_1 + Y_{32}a_2$, the total number of decision variables is $1+2+3 = 6$. For $\mathcal{U} = [0,10]^2$ (box on (a_1, a_2)) and a simple holding/shortage cost structure, solving the resulting 6-variable LP gives, e.g., $x_1^* = 5$, $Y_{21}^* = -0.5$ (order less in period 2 if period-1 demand was high, a natural mean-reversion adjustment), illustrating how the recourse matrix $\mathbf{Y}$ encodes interpretable feedback policies even though it was derived from a purely algebraic LP.

When affine rules are exact. ADR is not merely an approximation in every case: for problems where the true optimal recourse $\mathbf{y}^*(\mathbf{a})$ *is* affine – e.g. when the recourse problem has a unique non-degenerate optimal basis for *every* $\mathbf{a} \in \mathcal{U}$ – ADR recovers the *exact* two-stage robust optimum with zero approximation gap. This occurs whenever $\mathcal{U}$ is small enough (or the cost structure simple enough) that no basis change is triggered across $\mathcal{U}$; the suboptimal-

ity illustrated earlier arises precisely because the true recourse $\max(0, a-1)$ *does* undergo a kink (basis change) within $\mathcal{U} = [0, 2]$.

Finite adaptability and connections. An alternative to continuous decision rules is *finite adaptability*: partition $\mathcal{U}$ into K regions $\mathcal{U}_1, \ldots, \mathcal{U}_K$ and choose a separate $\mathbf{y}^{(k)}$ for each region, recovering the two-stage problem's flexibility in the limit $K \to \infty$ while keeping each sub-problem static-robust. The choice between ADR (smooth, $O(n)$ extra variables, polynomial-time) and finite adaptability (piecewise, $O(K)$ sub-problems, exponential in K for optimal partitions) mirrors the bias–variance and approximation-class trade-offs seen elsewhere in this textbook series (cf. Volume on optimization landscapes), and both serve as the deterministic backbone onto which the explicitly probabilistic multi-stage stochastic programs of Part 12 are built.

PART 10

Distributionally Robust Optimization

Optimization under ambiguity in the probability distribution itself, via moment-based and Wasserstein ambiguity sets.

Theorem: ADR feasibility is preserved under uncertainty-set inclusion

If $\mathcal{U}' \subseteq \mathcal{U}$ and $(\mathbf{x}, \mathbf{y}_0, \mathbf{Y})$ is feasible for the ADR-restricted problem with uncertainty set $\mathcal{U}$, then it is also feasible for $\mathcal{U}'$ (the same affine policy remains valid on any subset of the original uncertainty set). Consequently, the ADR-optimal objective for $\mathcal{U}'$ is *at least as good* (no worse) as the ADR-optimal objective for $\mathcal{U}$.

Proof. The robust constraints for $\mathcal{U}$ are $\sigma_{\mathcal{U}}(\cdot) \leq 0$ (in the notation of Theorem 9.5's proof); $\mathcal{U}' \subseteq \mathcal{U} \implies \sigma_{\mathcal{U}'}(\cdot) \leq \sigma_{\mathcal{U}}(\cdot)$ (support functions are monotone in set inclusion), so any $(\mathbf{x}, \mathbf{y}_0, \mathbf{Y})$ satisfying $\sigma_{\mathcal{U}}(\cdot) \leq 0$ also satisfies $\sigma_{\mathcal{U}'}(\cdot) \leq 0$. The objective-monotonicity follows since the feasible region for $\mathcal{U}'$ contains that for $\mathcal{U}$, so its minimum is $\leq$. ■

EXAMPLE — Quantifying the value of a tighter uncertainty set

Suppose $\mathcal{U} = [0, 2]$ (Section 9.5's worked example, ADR objective value $V(\mathcal{U})$) and a new sensor allows shrinking the demand-uncertainty interval to $\mathcal{U}' = [0.5, 1.5]$ (same midpoint, half the width). By the theorem, $V(\mathcal{U}') \leq V(\mathcal{U})$. Quantitatively, for the affine rule $y(a) = y_0 + Ya$ approximating $y^*(a) = \max(0, a-1)$: on $\mathcal{U} = [0, 2]$ the best affine rule had max error 0.25 (Section 9.5's worked example); on $\mathcal{U}' = [0.5, 1.5]$, fitting $y_0 + Ya$ to $\max(0, a-1)$ over the shorter interval $[0.5, 1.5]$ gives a *smaller* max error (the kink at $a = 1$ is now more "central" relative to a shorter interval, but the function is still non-affine over $[0.5, 1.5]$ unless the interval avoids the kink entirely – e.g. $\mathcal{U}'' = [1.2, 1.8] \subset (1, 2]$ avoids the kink, where $y^*(a) = a - 1$ is *exactly* affine, giving ADR error $= 0$). This illustrates concretely how investment in uncertainty reduction (better sensors, tighter forecasts) translates, via the monotonicity theorem, directly into ADR objective improvements – and, in the best case (uncertainty set avoiding all kinks of y^*), into *exact* (zero-gap) ADR.

Remarks and extensions. This section closes Part 9's progression from static (Sections 9.1–9.4) to dynamic (Section 9.5) robust optimization. A natural question, taken up properly in Part 12, is what happens when the "worst case over $\mathcal{U}$" of the static/adjustable RO framework is replaced by an "expectation over P" – the resulting two-stage *stochastic* program (Section 12.1) has an formally analogous adjustable structure (here-and-now $\mathbf{x}$, wait-and-see $\mathbf{y}(\boldsymbol{\xi})$), but optimizes $\mathbb{E}_{\boldsymbol{\xi}}[\cdot]$ rather than $\max_{\mathbf{a} \in \mathcal{U}}[\cdot]$. The affine decision rule of this section

and the value-function convexity of Section 12.1 are, in fact, dual perspectives on the same recourse structure – ADR *restricts* the recourse function to a tractable class, while Section 12.1's convexity result shows the *exact* (unrestricted) recourse function is already convex (hence, in principle, well-approximated by affine functions when the convex function is not too curved).

A final remark: the suboptimality bound for ADR (illustrated numerically in the worked example) can be tightened using *piecewise*-affine (rather than globally affine) decision rules, partitioning $\mathcal{U}$ into a small number of regions and using a separate affine rule on each – a hybrid between pure ADR and the finite-adaptability approach discussed at the end of this section, often striking a good balance between the $O(n)$ tractability of ADR and the exactness of fully general (but generally intractable) $\mathbf{y}(\cdot)$.

Theorem: ADR and linear-quadratic-Gaussian (LQG) control

For a linear dynamical system $\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t + \mathbf{w}_t$ with $\mathbf{w}_t \in \mathcal{U}_t$ and quadratic stage costs, the multi-stage affine-decision-rule policy $\mathbf{u}_t = \mathbf{u}_t^0 + \sum_{\tau<t} \mathbf{K}_{t,\tau}\mathbf{w}_\tau$ is equivalent to the linear state-feedback policy $\mathbf{u}_t = \mathbf{K}_t\mathbf{x}_t$ of LQG/LQR control theory, because $\mathbf{x}_t$ is itself an affine function of $(\mathbf{w}_0, \dots, \mathbf{w}_{t-1})$ for linear dynamics – "affine in disturbance history" and "linear in current state" are equivalent parametrizations.

Proof sketch. By the dynamics, $\mathbf{x}_t = \mathbf{A}^t\mathbf{x}_0 + \sum_{\tau<t} \mathbf{A}^{t-1-\tau}(\mathbf{B}\mathbf{u}_\tau + \mathbf{w}_\tau)$, an affine function of $(\mathbf{w}_0, \dots, \mathbf{w}_{t-1})$ given (by induction) affine policies $\mathbf{u}_\tau$. Conversely, any affine-in-history policy can be re-expressed, via this relation, as affine-in-$\mathbf{x}_t$, with $\mathbf{K}_t$ determined by the $\mathbf{K}_{t,\tau}$ via the state recursion. ■

Example — Reading an LQR gain as an ADR recourse matrix

A scalar LQR problem with $A = B = Q = R = 1$ has steady-state feedback gain $K = -(\sqrt{5}-1)/2 \approx -0.618$ (the negative reciprocal golden ratio, from the algebraic Riccati equation). Via the theorem, this $K \approx -0.618$ is *the same number* as an ADR recourse coefficient $Y_{t,t-1}$ in the equivalent multi-stage-ADR formulation – the $Y_{21} = 0.6$ "pass-through" coefficient of Section 9.5's inventory example is not merely *analogous* to an LQR gain, it *is* the same object under the affine-policy/state-feedback equivalence (for appropriately matched problem parameters) – a unification between robust optimization (1990s-2000s) and optimal control (1950s-60s) often obscured by differing notations.

Historical notes. Adjustable robust optimization was introduced by Ben-Tal, Goryashko, Guslitzer, and Nemirovski (2004, "Adjustable Robust Solutions of Uncertain Linear Programs"), with the affine decision rule (ADR) as the central tractable approximation. The terminology "decision rule," borrowed from stochastic control and dynamic programming (where a "policy" or "decision rule" is a function from observed history to action), signals the deep kinship between adjustable RO and the stochastic programming with recourse of Part 12 – a kinship made explicit in this section's Remarks. Finite adaptability (the K-region partition alternative to ADR) was developed somewhat later (Bertsimas & Caramanis, 2010) and remains an active research area, particularly regarding *how* to optimally choose the partition $\{\mathcal{U}_k\}_{k=1}^K$ – a genuinely hard combinatorial sub-problem for which only heuristic and approximation-guarantee results are currently known.

Theorem: exact optimality of affine decision rules under scalar uncertainty

For a two-stage adjustable RO problem with a *single* uncertain parameter $\xi \in [\underline{\xi}, \bar{\xi}] \subset \mathbb{R}$ entering the recourse problem's data affinely, and a recourse cost function $Q(\mathbf{x}, y, \xi)$ that is convex and piecewise-linear in ξ for every fixed $(\mathbf{x}, y)$ (as holds for any LP-representable recourse), the optimal adjustable-RO recourse policy $y^\star(\xi)$ is affine in ξ – the affine decision rule incurs *no* suboptimality gap whenever the uncertainty is genuinely one-dimensional.

Proof sketch. The worst case in a one-dimensional adjustable RO problem with convex (piecewise-linear) recourse cost in ξ is always attained at one of the two interval endpoints $\underline{\xi}, \bar{\xi}$ (a convex function on an interval attains its maximum at an extreme point of the interval); the optimal policy can therefore be characterized completely by its values $y(\underline{\xi})$ and $y(\bar{\xi})$ at the two endpoints, and the unique affine function through these two points, $y(\xi) = y(\underline{\xi}) + \frac{y(\bar{\xi}) - y(\underline{\xi})}{\bar{\xi} - \underline{\xi}}(\xi - \underline{\xi})$, matches the optimal recourse exactly at both endpoints and, by convexity of the recourse cost in ξ, performs at least as well as any other interpolation at every interior ξ – hence is itself optimal. This argument is special to one dimension (an interval has only two extreme points); in $\mathbb{R}^n$ for $n \geq 2$, a polytope's extreme points generally number far more than two and a single affine function cannot in general match an optimal policy's value at every extreme point simultaneously, so the argument does not extend. ■

Example — A case where the affine rule is exactly optimal

Two-stage inventory problem: first-stage order x, second-stage adjustment $y(\xi) = y^0 + \beta\xi$ in response to demand shock $\xi \in [-1, 1]$ (a single scalar uncertain parameter). The recourse cost is piecewise-linear and convex in ξ for fixed x, y^0, β; by the theorem, the affine rule $y(\xi) = y^0 + \beta\xi$ exactly attains the adjustable-RO optimum – no fully general (non-affine) recourse policy can improve on it, since the optimal policy's values at the two endpoints $\xi = \pm 1$ already determine, via the unique affine interpolant, a policy that is optimal everywhere on $[-1, 1]$ by the convexity argument above. This explains why the worked single-uncertain-parameter examples throughout Sections 9.4–9.5 always found affine recourse to be exactly optimal: with a genuinely scalar uncertain parameter, the theorem's hypothesis is automatically satisfied.

When the gap is strictly positive. For $n \geq 2$ uncertain parameters jointly constrained in a set $\mathcal{U} \subset \mathbb{R}^n$ (e.g. a high-dimensional ellipsoid or box), the one-dimensional argument above no longer applies, and the affine decision rule is generally strictly suboptimal: Bertsimas, Iancu, and Parrilo (2010, "Optimality of Affine Policies in Multistage Robust Optimization") show that for the ellipsoidal case in $\mathbb{R}^n$, the worst-case relative suboptimality of the affine rule is $O(\sqrt{n})$ in the dimension – growing, but only polynomially, which is the formal justification for affine rules' continued practical use even outside the scalar-uncertainty case covered by the theorem above: a bounded (if non-zero) suboptimality is, in most applications, an acceptable price for the dramatic ($\exp(n) \to \text{poly}(n)$) tractability gain.

Historical notes (addendum). The exact-optimality result for scalar uncertainty stated here complements, rather than supersedes, the general $O(\sqrt{n})$ suboptimality bound of Bertsimas–Iancu–Parrilo (2010): it identifies the simplest possible special case (a single uncertain parameter) in which the celebrated affine-rule tractability device of Ben-Tal et al. (2004) suffers *no* optimality loss whatsoever, while making transparent exactly why this convenient exactness fails to extend beyond one dimension – the endpoint-interpolation argument has no direct analogue once the uncertainty set's extreme points number more than two.

10.1 Ambiguity Sets: Moment-Based and ϕ-Divergence Balls

Overview

Distributionally robust optimization (DRO) replaces a single assumed distribution P_0 for a random vector $\boldsymbol{\xi}$ by an *ambiguity set* $\mathcal{P}$ of plausible distributions, and optimizes the worst case over $\mathcal{P}$: $\min_{\mathbf{x}} \sup_{P \in \mathcal{P}} \mathbb{E}_{\boldsymbol{\xi} \sim P}[h(\mathbf{x}, \boldsymbol{\xi})]$. This section introduces the two dominant families of ambiguity sets.

Example — Case study: ambiguity aversion in a pandemic policy decision

Early in a novel pathogen's emergence, policymakers face a choice between Intervention A (well-understood, moderate efficacy against *known* pathogen characteristics) and Intervention B (potentially higher efficacy, but its performance depends on *as-yet-unknown* viral characteristics, e.g. transmissibility mechanisms not yet characterised). Even if a (possibly optimistic) point-estimate expected efficacy of B exceeds A's known efficacy, ambiguity-averse policymakers may prefer A – directly analogous to preferring the known-probability red urn over the ambiguous black/yellow urn in Ellsberg's experiment. Crucially, Theorem 11.1 shows this preference is *not* reducible to "policymakers secretly believe B's true efficacy is lower than its point estimate" (that would be a SEU-consistent belief, not ambiguity aversion) – it is a structurally distinct attitude toward the *unknown-unknown* nature of B's efficacy distribution itself, requiring the Section 11.3–11.5 frameworks (not merely a pessimistic point-estimate adjustment) to formalize.

Moment-based ambiguity sets. Given estimates of the mean $\boldsymbol{\mu}_0$ and covariance $\square_0$ of $\boldsymbol{\xi}$, define

$$\mathcal{P}_{\text{mom}} = \left\{ P : \mathbb{E}_P[\boldsymbol{\xi}] = \boldsymbol{\mu}_0,\ \mathbb{E}_P[(\boldsymbol{\xi} - \boldsymbol{\mu}_0)(\boldsymbol{\xi} - \boldsymbol{\mu}_0)^\top] \preceq \square_0 \right\}$$

(or with equality $= \square_0$ in the stricter variant). This set contains *all* distributions with the specified first two moments (or moment bound), regardless of shape – including highly non-Gaussian, multi-modal, or discrete distributions.

Definition: ϕ-divergence

For a convex function $\phi : \mathbb{R}_{>0} \to \mathbb{R}$ with $\phi(1) = 0$, the ϕ-divergence between distributions P and Q (with P absolutely continuous w.r.t. Q) is

$$D_\phi(P \| Q) = \int \phi\left(\frac{dP}{dQ}\right) dQ.$$

Special cases: $\phi(t) = t \log t$ gives the KL divergence $D_{\text{KL}}(P \,\|\, Q)$; $\phi(t) = (t-1)^2$ gives the χ^2-divergence; $\phi(t) = \frac{1}{2}|t-1|$ gives (twice) total variation.

Example — Case study: emergency resource allocation

A disaster-relief agency must allocate a fixed budget across three response strategies, with payoffs (lives saved, in hundreds) depending on the disaster's eventual severity (mild/moderate/severe): Strategy $A = (8, 5, 1)$, Strategy $B = (4, 5, 5)$, Strategy $C = (2, 4, 7)$. **Maximin**: row minima $1, 4, 2$; max $= 4$, choose B. **Minimax regret**: column maxima $(8, 5, 7)$; regrets $A = (0, 0, 6)$, $B = (4, 0, 2)$, $C = (6, 1, 0)$; max regrets $6, 4, 6$; choose B (regret 4). **Hurwicz** $\alpha = 0.3$ (mildly optimistic, reflecting some confidence

that severe scenarios are less likely): scores $0.3\min + 0.7\max$: $A = 0.3(1) + 0.7(8) = 5.9$, $B = 0.3(4) + 0.7(5) = 4.7$, $C = 0.3(2) + 0.7(7) = 5.5$; choose A. The three criteria disagree (B, B, A) – a realistic outcome for genuinely difficult allocation decisions, and the disagreement itself is informative: it signals that the choice is sensitive to the decision-maker's attitude toward the (currently unknown) severity distribution, a sensitivity that the Γ-minimax and Choquet frameworks of Sections 11.3–11.4 can characterize more precisely than any single classical criterion.

ϕ-divergence balls. Given a reference (e.g. empirical) distribution $\hat{P}_N$ and radius $\epsilon > 0$,

$$\mathcal{P}_\phi(\epsilon) = \{P : D_\phi(P\|\hat{P}_N) \le \epsilon\}.$$

Unlike moment-based sets, ϕ-divergence balls are centred on a specific reference distribution and shrink to $\{\hat{P}_N\}$ as $\epsilon \to 0$, making them natural for data-driven settings (Section 10.4) where $\hat{P}_N$ is the empirical distribution of N samples.

Example — Comparing ambiguity sets on a discrete example

Let $\boldsymbol{\xi} \in \{1, 2, 3\}$ with empirical distribution $\hat{P}_N = (0.5, 0.3, 0.2)$, sample mean $\hat{\mu} = 1\cdot 0.5 + 2\cdot 0.3 + 3\cdot 0.2 = 1.7$, sample variance $\hat{\sigma}^2 = 0.5(1-1.7)^2 + 0.3(2-1.7)^2 + 0.2(3-1.7)^2 = 0.5(0.49) + 0.3(0.09) + 0.2(1.69) = 0.245 + 0.027 + 0.338 = 0.61$. The moment-based set $\mathcal{P}_{\text{mom}}$ with $\mu_0 = 1.7, \sigma_0^2 = 0.61$ includes, e.g., the two-point distribution P' placing mass 0.39 at $\xi = 1$ and 0.61 at $\xi = 3$ (which has the same mean and variance but is bimodal, very different from $\hat{P}_N$). In contrast, a KL-ball $\mathcal{P}_{\text{KL}}(\epsilon)$ for small ϵ only contains distributions *close in shape* to $\hat{P}_N$, excluding P' even though it shares the same first two moments – the two ambiguity sets thus protect against qualitatively different kinds of model misspecification.

The KL ambiguity set in closed form. For the KL-divergence ball $\mathcal{P}_{\text{KL}}(\epsilon) = \{P : D_{\text{KL}}\big(P \,\|\, \hat{P}_N\big) \le \epsilon\}$, the worst-case expectation admits the dual form (a special case of the Section 10.3 conic duality)

$$\sup_{P \in \mathcal{P}_{\text{KL}}(\epsilon)} \mathbb{E}_P[\ell(\boldsymbol{\xi})] = \inf_{\eta > 0} \left\{ \eta \log \mathbb{E}_{\hat{P}_N}\left[e^{\ell(\boldsymbol{\xi})/\eta}\right] + \eta\epsilon \right\}.$$

This log-sum-exp form is the same functional that appears as the cumulant generating function in large-deviations theory (Volume V) and as the softmax risk measure in reinforcement learning.

Example — KL-DRO on a finite outcome space

$\hat{P}_N = (0.5, 0.3, 0.2)$ on $\Xi = \{1, 2, 3\}$, $\ell(\xi) = \xi$, $\epsilon = 0.05$. Evaluate $g(\eta) = \eta \log(0.5e^{1/\eta} + 0.3e^{2/\eta} + 0.2e^{3/\eta}) + 0.05\eta$ at $\eta = 2$: $0.5e^{0.5} + 0.3e^{1} + 0.2e^{1.5} \approx 0.8244 + 0.8155 + 0.8963 = 2.5362$; $\log(2.5362) \approx 0.9305$; $g(2) \approx 2(0.9305) + 0.1 = 1.961$. Compare $\mathbb{E}_{\hat{P}_N}[\xi] = 1.7$. The gap 0.261 is the KL-robustness premium at $\eta = 2$; minimizing over $\eta > 0$ (numerically $\eta^* \approx 2.9$) gives the tightest bound, always ≥ 1.7 for $\epsilon > 0$.

Total variation as the bridging case. Total variation $D_{\text{TV}}(P, Q) = \sup_A |P(A) - Q(A)|$ corresponds to $\phi(t) = \frac{1}{2}|t - 1|$ and has the simple formula $\sup_{D_{\text{TV}}(P,\hat{P}_N) \le \epsilon} \mathbb{E}_P[\ell] = \mathbb{E}_{\hat{P}_N}[\ell] + \epsilon(\operatorname{ess\,sup} \ell - \mathbb{E}_{\hat{P}_N}[\ell])$ for $\epsilon \le 1$. TV-balls are exactly the ϵ-contamination neighbourhoods of Section 11.4, foreshadowing the decision-theoretic re-appearance of this same geometry.

Discussion. The choice between moment-based and divergence-based ambiguity sets reflects a bias toward, respectively, *distributional shape-agnostic* robustness (any distribution

with given moments, however exotic) versus *neighbourhood-of-empirical* robustness (distributions statistically close to observed data). Moment-based sets typically yield tractable DRO problems via semidefinite or conic duality (a generalisation of the conic arguments of Section 9.3), while ϕ-divergence and, as developed next in Section 10.2, Wasserstein balls connect naturally to the statistical theory of empirical processes and provide finite-sample coverage guarantees – the bridge between the geometric robust optimization of Part 9 and the genuinely probabilistic frameworks of Parts 11–12.

Theorem: moment ambiguity sets are invariant under affine transformations

If $\boldsymbol{\xi} \in \mathbb{R}^n$ has $\mathcal{P}_{\text{mom}} = \{P : \mathbb{E}_P[\boldsymbol{\xi}] = \boldsymbol{\mu}_0, \text{Cov}_P(\boldsymbol{\xi}) \preceq \square_0\}$, and $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\xi} + \mathbf{b}$ for fixed $\mathbf{A}, \mathbf{b}$, then the pushforward ambiguity set $\mathcal{P}'_{\text{mom}} = \{P' : \mathbb{E}_{P'}[\boldsymbol{\eta}] = \mathbf{A}\boldsymbol{\mu}_0 + \mathbf{b}, \text{Cov}_{P'}(\boldsymbol{\eta}) \preceq \mathbf{A}\square_0\mathbf{A}^\top\}$ exactly equals $\{P \circ (\mathbf{A} \cdot + \mathbf{b})^{-1} : P \in \mathcal{P}_{\text{mom}}\}$ – moment-based ambiguity "commutes" with affine transformations of the random vector.

Proof. Mean and covariance transform exactly as $\mathbb{E}[\mathbf{A}\boldsymbol{\xi} + \mathbf{b}] = \mathbf{A}\mathbb{E}[\boldsymbol{\xi}] + \mathbf{b}$ and $\text{Cov}(\mathbf{A}\boldsymbol{\xi} + \mathbf{b}) = \mathbf{A}\,\text{Cov}(\boldsymbol{\xi})\,\mathbf{A}^\top$ (bilinearity of covariance), and $\square \preceq \square_0 \implies \mathbf{A}\square\mathbf{A}^\top \preceq \mathbf{A}\square_0\mathbf{A}^\top$ (congruence preserves the PSD order). Both inclusions ($\subseteq$ and $\supseteq$) follow from these two facts applied to P and its pushforward. ■

Example — Currency conversion of a moment-ambiguous loss

A loss $\boldsymbol{\xi}$ (in EUR) has $\mathcal{P}_{\text{mom}}$ with $\mu_0 = 100$, $\sigma_0^2 = 400$ ($\sigma_0 = 20$). Converting to USD via $\eta = 1.1\xi$ (exchange rate, treated as fixed for this exercise): by the theorem, $\mathcal{P}'_{\text{mom}}$ has $\mu_0' = 110$, $\sigma_0'^2 = 1.21 \times 400 = 484$ ($\sigma_0' = 22$). One can verify directly: if P has mean 100, variance ≤ 400, then $\eta = 1.1\xi$ has mean $1.1 \times 100 = 110$ and variance $1.21 \times \text{Var}(\xi) \leq 1.21 \times 400 = 484$ ✓. This "commutativity" means moment-DRO models can be built in whatever units (or, more generally, whatever *linear transformation* of the raw data – e.g. standardized/whitened features) are most convenient, and the ambiguity-set parameters transform predictably without needing to re-derive them from scratch in the new coordinates.

Remarks and extensions. The ϕ-divergence family is considerably richer than the three special cases (KL, χ^2, TV) highlighted in this section. The *Cressie-Read* family $\phi_\gamma(t) = \frac{t^\gamma - \gamma t + \gamma - 1}{\gamma(\gamma-1)}$ for $\gamma \in \mathbb{R} \setminus \{0, 1\}$ interpolates continuously between KL ($\gamma \to 1$), reverse-KL ($\gamma \to 0$), and χ^2 ($\gamma = 2$), and the corresponding DRO dual (a generalisation of the KL dual of Section 10.1's first addition) has the form $\inf_{\eta>0}\{\eta\,\mathbb{E}_{\hat{P}_N}[\phi_\gamma^*(\ell/\eta)] + \eta\epsilon\}$ for the convex conjugate ϕ_γ^* – a single unifying formula that specializes, for each γ, to give a different "risk measure" on the empirical loss distribution. This Cressie-Read family is widely used in empirical likelihood methods in statistics, an entirely separate literature that turns out to share, via this ϕ-divergence connection, much of its mathematical core with DRO.

A second remark: although moment-based and ϕ-divergence ambiguity sets are presented here as alternatives, they can be *combined* – e.g. $\mathcal{P} = \{P : D_\phi(P\|\hat{P}_N) \leq \epsilon,\ \mathbb{E}_P[\boldsymbol{\xi}] = \boldsymbol{\mu}_0\}$ – giving ambiguity sets that are simultaneously "close to the data" (small divergence) and "consistent with known structure" (matching moments). The dual of such combined ambiguity sets is correspondingly a combination of the two dual penalties (a divergence-based term plus a moment-matching Lagrange term), and remains tractable whenever each piece is.

Theorem: moment ambiguity sets and the Gaussian maximum-entropy connection

Among all distributions P with $\mathbb{E}_P[\boldsymbol{\xi}] = \boldsymbol{\mu}_0$ and $\mathrm{Cov}_P(\boldsymbol{\xi}) = \square_0$ (equality), the multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_0, \square_0)$ *uniquely maximizes* differential entropy $H(P) = -\int p \log p$. Consequently $\mathcal{P}_{\text{mom}}$ can be seen as "centred" on the Gaussian information-theoretically, even though $\mathcal{P}_{\text{mom}}$ contains arbitrarily non-Gaussian P.

Proof sketch. A Lagrangian argument: maximize $H(P)$ subject to $\int p = 1, \int \boldsymbol{\xi} p = \boldsymbol{\mu}_0, \int \boldsymbol{\xi}\boldsymbol{\xi}^\top p = \mathbf{M}_0$. The first-order condition gives $\log p(\boldsymbol{\xi}) = -1 - \lambda_0 - \boldsymbol{\lambda}_1^\top \boldsymbol{\xi} - \boldsymbol{\xi}^\top \square_2 \boldsymbol{\xi}$, i.e. $p \propto \exp(-\boldsymbol{\xi}^\top \square_2 \boldsymbol{\xi} - \boldsymbol{\lambda}_1^\top \boldsymbol{\xi})$, a Gaussian density with $(\boldsymbol{\lambda}_1, \square_2)$ fixed by matching $(\boldsymbol{\mu}_0, \square_0)$. ■

Example — Gaussian as reference, not worst case

For the credit-risk case study ($\hat{\mu} = 3\%, \hat{\sigma} = 1\%$), the max-entropy distribution is $\mathcal{N}(3\%, 1\%^2)$. But the moment-DRO *worst case* for a tail-probability objective is generally *not* Gaussian – typically a two-point distribution at moment-constrained extremes (as in the bimodal example of Section 10.1), because such distributions *maximize*, rather than typify, the moment-constrained class. The Gaussian's max-entropy status makes it a natural *default* when no further information is available, but moment-DRO's robustness guarantee comes precisely from *not* assuming this default is correct – the worst case lives at the *extreme points* of $\mathcal{P}_{\text{mom}}$, not its informational "centre."

Historical notes. Moment-based ambiguity (the "distribution-free" worst case over all P with given moments) has roots in the classical *Chebyshev-Markov* bounds of 19th-century probability theory, but its systematic application to optimization – the moment problem as an SDP, Section 10.3 – is due to Scarf (1958) in the context of the newsvendor problem (a striking historical loop: Scarf's 1958 "min-max newsvendor" is, in modern terms, exactly a moment-based DRO instance of the very newsvendor problem used as this volume's running pedagogical example in Section 12.1 and 12.5). ϕ-divergence ambiguity sets emerged later from the *robust statistics* and *empirical likelihood* literatures (Owen, 1988, and others), with their optimization application crystallising in the 2000s-2010s (Ben-Tal et al., 2013, "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities"). The 2010s also saw the moment-based and ϕ-divergence strands increasingly cross-referenced with the (initially separate) Wasserstein-DRO literature discussed in Section 10.2's notes, converging toward the unified "DRO" terminology and framework used throughout Part 10.

Theorem: nesting of ϕ-divergence ambiguity sets across divergence families

For two ϕ-divergences ϕ_1, ϕ_2 with $D_{\phi_1}(Q\|P) \le D_{\phi_2}(Q\|P)$ for every pair of distributions Q, P on the relevant space (e.g. $\phi_1 = \mathrm{KL}$ and $\phi_2 = \chi^2$, which satisfy $D_{\mathrm{KL}}(Q\|P) \le D_{\chi^2}(Q\|P)$ pointwise – a classical inequality obtained by applying Jensen's inequality to the concave function log in the form $D_{\mathrm{KL}}(Q\|P) \le \log(1 + D_{\chi^2}(Q\|P)) \le D_{\chi^2}(Q\|P)$, the last step using $\log(1+x) \le x$), the corresponding ambiguity sets satisfy $\mathcal{P}_{\phi_1}(\eta) \supseteq \mathcal{P}_{\phi_2}(\eta)$ for the same radius η – a pointwise-smaller divergence function yields a *larger* (less restrictive) ambiguity ball at fixed radius.

Proof. If $D_{\phi_1}(Q\|\hat{P}_N) \le D_{\phi_2}(Q\|\hat{P}_N)$ for every Q, then $D_{\phi_2}(Q\|\hat{P}_N) \le \eta \implies D_{\phi_1}(Q\|\hat{P}_N) \le \eta$ directly, i.e. every Q in the ϕ_2-ball also lies in the ϕ_1-ball, giving $\mathcal{P}_{\phi_2}(\eta) \subseteq \mathcal{P}_{\phi_1}(\eta)$. ■

Example — Quantifying the KL-vs-χ^2 gap on a concrete pair

For $P = (0.4, 0.6)$, $Q = (0.5, 0.5)$ (two-point distributions): $D_{\mathrm{KL}}(Q\|P) = 0.5\ln(0.5/0.4) + 0.5\ln(0.5/0.6) = 0.5(0.223) + 0.5(-0.182) = 0.0203$. $D_{\chi^2}(Q\|P) = \sum_i (Q_i - P_i)^2/P_i = (0.1)^2/0.4 + (-0.1)^2/0.6 = 0.025 + 0.0167 = 0.0417$. Indeed $0.0203 < 0.0417$, confirming the pointwise ordering for this pair; consequently a KL-ball of radius $\eta = 0.03$ around P *contains* Q (since $0.0203 < 0.03$), while a χ^2-ball of the *same* radius 0.03 does *not* contain Q (since $0.0417 > 0.03$) – the same nominal radius parameter η encodes a materially different (here, roughly half as permissive) ambiguity set depending on which divergence family is chosen, a calibration subtlety often overlooked when practitioners select η by analogy across different DRO formulations.

Practical calibration implication. Because the nesting theorem shows that divergence *choice* and divergence *radius* jointly determine the effective ambiguity set, a robust practice is to calibrate η *within* a fixed divergence family using the sample-complexity tools of Section 10.4 (which already account for the specific ϕ chosen), rather than attempting to transfer a calibrated η from one ϕ-divergence family to another by analogy – the nesting theorem shows such transfers can be off by a divergence-dependent multiplicative factor that is generally distribution-dependent (as the worked example illustrates) rather than a fixed universal constant.

Historical notes (addendum). The systematic comparison of ϕ-divergence families by their pointwise ordering, while a classical fact in information theory (going back to Csiszár's 1967 axiomatic treatment of f-divergences), was first connected explicitly to DRO ambiguity-set calibration by Ben-Tal, den Hertog, De Waegenaere, Melenberg, and Rennen (2013, "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities"), whose systematic comparison table across ϕ-divergence families remains the standard practical reference for divergence selection in ϕ-divergence DRO.

10.2 Wasserstein Distributionally Robust Optimization

Overview

The Wasserstein distance $W_p(P, Q)$ measures the minimal "transportation cost" to morph distribution P into Q. Wasserstein DRO (WDRO) defines $\mathcal{P} = \{P : W_p(P, \hat{P}_N) \leq \epsilon\}$, and – remarkably – admits a tractable dual reformulation as a regularized empirical risk minimization, directly connecting robustness to regularization (cf. Section 10.5).

Example — Case study: expert elicitation via a capacity

A panel of three climate experts gives subjective probabilities for "global temperature rise exceeds 2°C by 2100": Expert 1 says 60%, Expert 2 says 75%, Expert 3 says 50%. Rather than averaging (giving 61.7%, discarding disagreement information), a capacity ν can be constructed with $\nu(\{> 2°C\}) = \min(60, 75, 50)\% = 50\%$ (pessimistic/subadditive aggregation) and $\nu(\{\leq 2°C\}) = \min(40, 25, 50)\% = 25\%$ – note $\nu(\{> 2°C\}) + \nu(\{\leq 2°C\}) = 75\% < 100\%$, the hallmark of subadditivity, reflecting that the experts' *disagreement itself* is treated as a source of ambiguity (neither outcome gets the "benefit of the doubt"). The Choquet integral of a policy's payoff (which depends on whether 2°C is exceeded) with respect to this ν would then weight the "bad news exceeds 2°C" scenario more heavily than any individual expert's probability, formal-

izing a precautionary aggregation of expert disagreement – an application of Section 11.3's machinery to expert-elicitation problems common in climate-policy and other long-horizon risk assessments.

Definition: Wasserstein distance

For $p \geq 1$ and distributions P, Q on a metric space (Ξ, d), the p-Wasserstein distance is

$$W_p(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}')^p \, d\pi(\boldsymbol{\xi}, \boldsymbol{\xi}') \right)^{1/p},$$

where $\Pi(P,Q)$ is the set of all joint distributions ("couplings") with marginals P and Q. Intuitively, π specifies how much "mass" is moved from each $\boldsymbol{\xi}$ under P to each $\boldsymbol{\xi}'$ under Q, and W_p is the minimal p-th-root-averaged transport cost $d(\boldsymbol{\xi}, \boldsymbol{\xi}')$.

Example — Case study: robust hiring decisions under demographic prior uncertainty

A firm calibrates a hiring-screening algorithm's threshold using historical applicant-outcome data, yielding a prior π_0 over (qualification level, demographic group) pairs. Concerned that π_0 may not represent the *future* applicant pool (e.g. due to outreach-program-driven demographic shifts), the firm adopts $\Gamma = \{(1-\epsilon)\pi_0 + \epsilon Q : Q \text{ arbitrary}\}$ for $\epsilon = 0.15$ (anticipating up to 15% pool composition shift). The Γ-minimax threshold (minimizing worst-case mis-screening cost over Γ) will, by the cutting-plane procedure of Section 11.4, generally be *more conservative* (e.g. a lower bar, admitting more candidates for further review) than the π_0-Bayes-optimal threshold – directly analogous to the a_2-over-a_1 switch in the chapter's running example, but now with the "contamination" Q representing genuinely plausible future applicant-pool compositions rather than an abstract adversary. This case illustrates Γ-minimax's applicability to fairness-adjacent robustness questions: protecting a decision rule against *distributional drift* in the population it will be applied to, a concern of major practical importance for any data-driven screening system.

The WDRO problem and its dual. For loss function $\ell(\mathbf{x}, \boldsymbol{\xi})$, the WDRO problem is

$$\min_{\mathbf{x}} \sup_{P : W_p(P, \hat{P}_N) \leq \epsilon} \mathbb{E}_{\boldsymbol{\xi} \sim P}[\ell(\mathbf{x}, \boldsymbol{\xi})].$$

Theorem: Strong duality for WDRO (Mohajerin Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019)

For $p = 1$ and $\ell(\mathbf{x}, \cdot)$ Lipschitz with constant $L(\mathbf{x})$ w.r.t. d,

$$\sup_{P : W_1(P, \hat{P}_N) \leq \epsilon} \mathbb{E}_P[\ell(\mathbf{x}, \boldsymbol{\xi})] = \mathbb{E}_{\hat{P}_N}[\ell(\mathbf{x}, \boldsymbol{\xi})] + \epsilon \, L(\mathbf{x}),$$

so the WDRO problem becomes $\min_{\mathbf{x}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}, \hat{\boldsymbol{\xi}}_i) + \epsilon \, L(\mathbf{x})$ – the empirical risk plus a *Lipschitz-constant penalty*, an explicit regularization term.

Example — Case study: bridge design under load-uncertainty

A bridge's design load capacity $\mathbf{c}(\mathbf{x})^\top \mathbf{u}_0$ (a function of design choices $\mathbf{x}$: girder dimensions, material grade, etc.) must exceed a code-mandated minimum r_c. Design $\mathbf{x}_A$

exactly meets r_c at the nominal (best-estimate) load conditions $\mathbf{u}_0$ using minimal material (lowest cost). Design $\mathbf{x}_B$ exceeds r_c by a margin, using 15% more material (higher cost). By Theorem 11.5, $\hat{\alpha}(\mathbf{x}_A, r_c) = 0$ (zero robustness to any load increase beyond nominal – e.g. from unanticipated traffic growth, or updated seismic hazard maps), while $\hat{\alpha}(\mathbf{x}_B, r_c) > 0$ (some tolerance for load increases). Civil engineering codes' mandated "safety factors" (typically requiring design capacity to exceed nominal load by 50-100% for many structural elements) are, in this light, *regulatorily-mandated minimum info-gap robustness levels* $\hat{\alpha}_{\min}$ – Theorem 11.5's closed form shows that, for linear capacity models, this translates directly into a minimum required value of $\|\mathbf{c}(\mathbf{x})\|_*$ relative to the slack $\mathbf{c}(\mathbf{x})^\top \mathbf{u}_0 - r_c$, giving structural engineers an explicit formula connecting a code-mandated safety factor to the info-gap framework of this section.

Proof sketch. The inner supremum, by Kantorovich–Rubinstein duality, equals $\mathbb{E}_{\hat{P}_N}[\ell(\mathbf{x}, \boldsymbol{\xi})] + \epsilon \sup_{\boldsymbol{\xi}'} \frac{\ell(\mathbf{x}, \boldsymbol{\xi}') - \ell(\mathbf{x}, \hat{\boldsymbol{\xi}})}{d(\boldsymbol{\xi}', \hat{\boldsymbol{\xi}})}$ in the limit, and the supremum of difference quotients is exactly the Lipschitz constant $L(\mathbf{x}) = \sup_{\boldsymbol{\xi} \neq \boldsymbol{\xi}'} \frac{|\ell(\mathbf{x}, \boldsymbol{\xi}) - \ell(\mathbf{x}, \boldsymbol{\xi}')|}{d(\boldsymbol{\xi}, \boldsymbol{\xi}')}$. A fully rigorous proof requires care with the measure-theoretic limit but the result is exact for $p = 1$ under mild regularity. ■

Example — Linear regression as a WDRO instance

Take $\ell(\mathbf{x}, (\mathbf{u}, v)) = |v - \mathbf{x}^\top \mathbf{u}|$ (absolute regression loss), $d((\mathbf{u}, v), (\mathbf{u}', v')) = \|\mathbf{u} - \mathbf{u}'\|_q + |v - v'|$ for dual norm exponent q (Hölder-conjugate to the norm appearing in ℓ). One can show $L(\mathbf{x}) = \|\mathbf{x}\|_{q^*} \vee 1$ (where q^* is conjugate to q), so the WDRO objective becomes $\frac{1}{N}\sum_i |v_i - \mathbf{x}^\top \mathbf{u}_i| + \epsilon \|\mathbf{x}\|_{q^*}$ – for $q^* = 2$ this is exactly *ridge-regularized* least-absolute-deviation regression, and for $q^* = \infty$ it is LASSO-like ℓ_1-regularized regression. The Wasserstein radius ϵ plays the role of the regularization strength λ, with the crucial difference that ϵ has a direct interpretation as a guaranteed-robustness radius in the data space.

Higher-order Wasserstein and the $p = 2$ case. For $p = 2$, $W_2(P, Q)^2 = \inf_\pi \int \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2^2 \, d\pi$. Unlike $p = 1$, the $p = 2$ dual does not reduce to a simple Lipschitz-penalty form for general ℓ; however, for ℓ *quadratic* in $\boldsymbol{\xi}$ (mean-variance portfolio problems), the W_2-DRO dual is a tractable SDP, connecting to the moment-SDP duality of Section 10.3.

Example — Verifying the $p = 1$ dual numerically

$\Xi = \mathbb{R}$, $\hat{P}_N = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2$, $\ell(\xi) = |\xi - x|$ at $x = 1$, $d(\xi, \xi') = |\xi - \xi'|$, $\epsilon = 0.3$. ℓ is 1-Lipschitz ($L = 1$), giving $\sup_{P: W_1 \leq 0.3} \mathbb{E}_P[|\xi - 1|] = \mathbb{E}_{\hat{P}_N}[|\xi - 1|] + 0.3$. $\mathbb{E}_{\hat{P}_N}[|\xi - 1|] = \frac{1}{2}(1) + \frac{1}{2}(1) = 1$. Dual value $= 1.3$. *Verification*: $P' = \frac{1}{2}\delta_{-0.3} + \frac{1}{2}\delta_{2.3}$ has $W_1(P', \hat{P}_N) \leq 0.3$ ✓, and $\mathbb{E}_{P'}[|\xi - 1|] = \frac{1}{2}(1.3) + \frac{1}{2}(1.3) = 1.3$ ✓ – the bound is attained.

Group-structured Wasserstein penalties. If $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_G)$ has a group structure and $d(\boldsymbol{\xi}, \boldsymbol{\xi}') = \sum_g w_g \|\boldsymbol{\xi}_g - \boldsymbol{\xi}_g'\|$, the WDRO dual penalty becomes a *group-norm* (group LASSO) regularizer with group-specific strengths ϵ / w_g – group weights encode differential trust in each data group.

Implications. This duality result formally unifies distributional robustness with the regularization techniques used throughout statistical learning (ridge, LASSO, and – as Section 10.5 will show – adversarial training penalties all emerge as special cases of a Wasserstein-robustness radius). It also explains *why* regularization improves out-of-sample (test) performance: a regularized estimator is, by this equivalence, simultaneously optimal for a whole

neighbourhood of the empirical distribution, including (with high probability, by the concentration results of Section 10.4) the true unknown data-generating distribution P^* itself.

Theorem: triangle inequality for Wasserstein distance and its DRO implication

W_p satisfies the triangle inequality: $W_p(P,R) \le W_p(P,Q) + W_p(Q,R)$ for any P, Q, R. Consequently, if $\hat{P}_N$ is the empirical distribution and P^* the true distribution, then for any P with $W_p(P, \hat{P}_N) \le \epsilon$,

$$W_p(P, P^*) \le \epsilon + W_p(\hat{P}_N, P^*).$$

Proof. The triangle inequality for W_p follows from the fact that W_p is a genuine metric on the space of probability measures with finite p-th moments (a consequence of the existence of "glued" couplings: given near-optimal couplings π_{PQ} and π_{QR}, one can construct a coupling of P and R via the conditional structure of Q, whose cost is bounded by $W_p(P,Q) + W_p(Q,R)$ via Minkowski's inequality for the L^p norm of transport costs). The displayed inequality is then the triangle inequality with $Q = \hat{P}_N$, $R = P^*$. ∎

Example — Combining WDRO and concentration bounds

By Theorem 10.4, $W_1(\hat{P}_N, P^*) \le \epsilon_N(\beta)$ with probability $\ge 1 - \beta$. Combined with the triangle inequality above, any P in the WDRO ball of radius ϵ around $\hat{P}_N$ satisfies $W_1(P, P^*) \le \epsilon + \epsilon_N(\beta)$ with probability $\ge 1 - \beta$ – i.e. the *true* distribution P^* is itself within $\epsilon_N(\beta)$ of the ball's boundary (not necessarily inside the ball, but close to it). If $\epsilon \ge \epsilon_N(\beta)$ exactly (the calibration of Section 10.4), then by the reverse triangle inequality $W_1(P^*, \hat{P}_N) \le \epsilon_N(\beta) \le \epsilon$, so P^* is *inside* the ball, recovering Section 10.4's coverage guarantee as a special case of this more general triangle-inequality statement – which additionally quantifies *how far outside* the ball P^* could be if $\epsilon < \epsilon_N(\beta)$ (under-calibrated WDRO), a useful sensitivity statement absent from the basic coverage guarantee.

Remarks and extensions. The Wasserstein-DRO duality of Theorem 10.2, proven here for $p = 1$, has analogues for general $p \ge 1$, but the dual generally involves an *infimal convolution* $(\ell \oplus (c \cdot d^p))(\boldsymbol{\xi}) = \inf_{\boldsymbol{\xi}'} \{\ell(\mathbf{x}, \boldsymbol{\xi}') + c\, d(\boldsymbol{\xi}, \boldsymbol{\xi}')^p\}$ rather than a simple additive Lipschitz penalty – the $p = 1$ case is special in that the infimal convolution with a (homogeneous-degree-1) norm-based cost reduces to the simple additive form. For $p > 1$, the dual problem is itself an optimization (the infimal convolution), so WDRO for $p > 1$ is, computationally, a *nested* optimization – still tractable for many ℓ (e.g. ℓ convex and d^p strongly convex makes the inner inf a well-behaved proximal-operator-like computation, familiar from the proximal-gradient methods covered in the optimization volumes of the companion series), but structurally more involved than the $p = 1$ case emphasised in this section.

A further important remark concerns the choice of *ground metric* d in applications: for tabular data with mixed continuous/categorical features, d is often taken as a weighted combination $d((\mathbf{u}, v), (\mathbf{u}', v')) = \|\mathbf{u} - \mathbf{u}'\|_2 + \lambda \cdot \mathbf{1}[v \ne v']$ (Euclidean on continuous features, Hamming-type on categorical), and the resulting WDRO dual penalty correspondingly combines an ℓ_2-norm regularizer (continuous features) with an effectively combinatorial penalty (categorical features) – an active area where the clean theory of this section meets considerable practical complexity.

Theorem: WDRO with combined feature/label perturbation costs

If $d((\mathbf{u}, v), (\mathbf{u}', v')) = \|\mathbf{u} - \mathbf{u}'\| + \kappa|v - v'|$ for binary labels and trade-off $\kappa > 0$, the WDRO dual decomposes additively: for ambiguity radius $\epsilon < \kappa$, label flips are "too expensive" and the dual reduces to the usual feature-perturbation regularizer $\epsilon \|\nabla_{\mathbf{u}} \ell\|_{q^*}$; for $\epsilon \geq \kappa$, both feature- and label-perturbation channels contribute to the dual.

Proof sketch. The Wasserstein ball of radius ϵ around $\hat{P}_N$ (with this d) contains feature-perturbed points (cost $\|\mathbf{u} - \mathbf{u}'\| \leq \epsilon$, $v' = v$) whenever $\epsilon \geq 0$, and label-flipped points (cost $\kappa|v - v'| = \kappa$) only if $\epsilon \geq \kappa$. The Kantorovich dual's sup over perturbation *types* for each transported mass unit therefore activates the label-flip channel only once $\epsilon \geq \kappa$. ■

Example — Calibrating κ from a known label-noise rate

If labels are flipped with probability $\eta = 2\%$, a natural calibration is $\kappa \approx \eta \times$ (loss increase from a flip). For 0-1 loss (flip changes ℓ from 0 to 1): $\kappa \approx 0.02 \times 1 = 0.02$. If the feature-perturbation radius $\epsilon_{\text{feat}} = 0.1 > \kappa = 0.02$ (chosen via Section 10.4's methods), the "$\epsilon \geq \kappa$" regime applies – the WDRO model automatically incorporates *both* feature- and label-noise-robustness, the latter governed by η via κ – a principled alternative to ad-hoc label-smoothing heuristics.

Historical notes. Wasserstein DRO's modern form – and especially the regularization-duality result of Theorem 10.2 – emerged essentially simultaneously from two independent research threads around 2017-2018: Mohajerin Esfahani and Kuhn (2018, building on earlier 2015 work) approached it from a stochastic-programming/operations-research perspective, while Blanchet, Kang, and Murthy (2019) and Sinha, Namkoong, and Duchi (2018) approached it from a statistics/machine-learning perspective (the latter explicitly framing it as a principled foundation for adversarial training, Section 10.5). This convergence – two communities independently rediscovering the same duality from different motivating applications – is itself evidence of the result's fundamental nature, and mirrors the Section 9.3 S-lemma's similarly independent rediscovery across control theory and optimization. The Wasserstein distance itself has a much longer history, dating to Monge's 1781 transportation problem and Kantorovich's 1942 relaxation (for which Kantorovich shared the 1975 Nobel Prize in Economics) – one of the longer historical arcs connecting a foundational mathematical idea to its eventual optimization application, spanning nearly two and a half centuries.

Theorem: cost-metric sensitivity of the Wasserstein ball's effective radius

For the Wasserstein-DRO dual penalty $\epsilon \|\mathbf{c}\|_{*,d}$ (Section 10.2/10.3's worked Lipschitz-penalty form), where $\|\cdot\|_{*,d}$ denotes the dual norm *induced by the chosen ground-cost metric* d, switching the ground cost from $d_1(\xi, \xi') = \|\xi - \xi'\|_1$ to $d_2(\xi, \xi') = \|\xi - \xi'\|_2$ in $\mathbb{R}^k$ changes the effective penalty by a factor bounded by $\sqrt{k}$ in either direction: $\|\mathbf{c}\|_\infty \leq \|\mathbf{c}\|_2 \leq \sqrt{k} \|\mathbf{c}\|_\infty$ (the dual norms of ℓ_1 and ℓ_2 ground costs being ℓ_∞ and ℓ_2 respectively).

Proof. Standard norm-equivalence inequalities in $\mathbb{R}^k$: $\|\mathbf{c}\|_\infty \leq \|\mathbf{c}\|_2 \leq \sqrt{k} \|\mathbf{c}\|_\infty$, applied to the dual norms that appear in the Kantorovich-dual penalty term for ℓ_1- and ℓ_2-ground-cost Wasserstein balls respectively. ■

Example — A concrete dimension-dependence illustration

For a $k = 50$-dimensional feature vector with gradient $\mathbf{c}$ having all 50 components equal to 1 (a maximally "spread" gradient), $\|\mathbf{c}\|_\infty = 1$ but $\|\mathbf{c}\|_2 = \sqrt{50} \approx 7.07$ – a $7\times$ difference purely from the choice of ground-cost metric, even though both the underlying data and the radius ϵ are unchanged. This means a Wasserstein-DRO model's regularization strength can swing by an order of magnitude simply by reparametrizing the ground cost from ℓ_2 to ℓ_1 (or a weighted variant), with no change to the data or to ϵ – the ground-cost metric is not a minor technical choice but a first-order modelling decision, on par with the choice of ϵ itself, and should be reported alongside ϵ whenever a Wasserstein-DRO model's regularization strength is being communicated or compared across studies.

A worked comparison with ϕ-divergence radii. This metric-sensitivity issue has no direct analogue in ϕ-divergence DRO (Section 10.1), whose ambiguity sets are intrinsically dimension-free (a χ^2-ball radius η means the same thing regardless of the dimension k of Ξ) – a genuine advantage of ϕ-divergence ambiguity sets for cross-study comparability, traded off against their well-known restriction to ambiguity sets supported on the same finite point-mass locations as $\hat{P}_N$ (Section 10.1's overview), a restriction the Wasserstein ball, with its associated metric-choice subtlety, does not share.

Historical notes (addendum). The ground-cost-metric sensitivity of Wasserstein-DRO penalties is well known among specialists (it is implicit in the very definition of the dual-norm penalty derived by Mohajerin Esfahani and Kuhn (2018) and Blanchet, Kang, and Murthy (2019)) but is frequently glossed over in applied DRO papers that report only ϵ without specifying the ground-cost metric and its induced dual norm – a reproducibility gap this section's explicit dimension-dependence bound is intended to help close.

10.3 Duality Results for Distributionally Robust Problems

Overview

This section develops the general semi-infinite-to-finite duality argument underlying both the moment-based and Wasserstein DRO reformulations, casting $\sup_{P\in\mathcal{P}} \mathbb{E}_P[\ell(\mathbf{x}, \boldsymbol{\xi})]$ as the value of an infinite-dimensional LP over measures, and dualizing it into a finite-dimensional problem over dual "price" functions/variables.

Example — Case study: vaccine-dose ordering under demand uncertainty

A public health agency orders vaccine doses x months before a flu season, with seasonal demand ξ historically $\sim \mathcal{N}(2{,}000{,}000, (400{,}000)^2)$. Overordering wastes doses at cost $h = \$5$/dose (storage/disposal); underordering incurs cost $p = \$40$/dose (emergency procurement + public health cost of unmet demand). The critical ratio is $(p - c)/(p + h)$ where $c \approx \$8$/dose (the order cost itself, distinct from over/under penalties) gives $(40 - 8)/(40 + 5) = 32/45 \approx 0.711$. The order quantity is the 71.1st percentile of demand: $x^* = 2{,}000{,}000 + 0.711$-quantile-$z \times 400{,}000$. For $\mathcal{N}$, the 71.1% quantile is $z \approx 0.555$, giving $x^* \approx 2{,}000{,}000 + 222{,}000 = 2{,}222{,}000$ doses – noticeably above the mean, reflecting the asymmetric cost structure ($p > h$, so the agency rationally "overshoots" the mean to avoid the costlier shortage outcome). This is a direct, high-stakes application of the newsvendor critical-ratio formula of Theorem 12.1 to a problem where the recourse cost asymmetry ($p/h = 8$) has genuine public-health

significance.

The primal moment problem. For the moment ambiguity set $\mathcal{P}_{\text{mom}} = \{P \succeq 0 : \int dP = 1,\ \int \boldsymbol{\xi}\, dP = \boldsymbol{\mu}_0,\ \int \boldsymbol{\xi}\boldsymbol{\xi}^\top dP \preceq \mathbf{M}_0\}$ (writing the covariance constraint via the second moment matrix $\mathbf{M}_0 = \square_0 + \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top$), the inner problem

$$\sup_{P \in \mathcal{P}_{\text{mom}}} \int \ell(\mathbf{x}, \boldsymbol{\xi})\, dP(\boldsymbol{\xi})$$

is a linear functional of P subject to linear (in P) moment-matching constraints – a generalized moment problem (GMP), an infinite-dimensional LP.

Theorem: Dual of the generalized moment problem

The GMP above has dual

$$\inf_{y_0,\, \mathbf{y}_1,\, \mathbf{Y}_2 \succeq 0} \ y_0 + \mathbf{y}_1^\top \boldsymbol{\mu}_0 + \operatorname{tr}(\mathbf{Y}_2 \mathbf{M}_0) \quad \text{s.t.} \quad y_0 + \mathbf{y}_1^\top \boldsymbol{\xi} + \boldsymbol{\xi}^\top \mathbf{Y}_2 \boldsymbol{\xi} \geq \ell(\mathbf{x}, \boldsymbol{\xi}) \quad \forall \boldsymbol{\xi}.$$

Under mild conditions (e.g. $\ell(\mathbf{x}, \cdot)$ piecewise linear or concave-representable), strong duality holds and the pointwise constraint "$\forall \boldsymbol{\xi}$" can itself be converted to finitely many SDP constraints via the S-lemma or sum-of-squares relaxations, yielding an overall SDP in $\mathbf{x}$ and the dual variables $(y_0, \mathbf{y}_1, \mathbf{Y}_2)$.

Example — Case study: reducing 10,000 weather scenarios to 20 for a power-grid model

A power-grid unit-commitment model uses historical weather data to generate $N = 10{,}000$ scenarios of (wind speed, temperature) pairs affecting both renewable generation and demand. Solving the resulting extensive-form LP with $10{,}000$ scenario-blocks is computationally prohibitive for the agency's daily re-planning cycle. Wasserstein-based reduction (Theorem 12.2) to $S' = 20$ scenarios, run daily, takes seconds (a small transportation LP) and yields a 20-scenario extensive form solvable in the required time window. Empirically, such reductions (a common practice in grid-operator software) typically retain 95-99% of the $10{,}000$-scenario solution's expected-cost performance when validated out-of-sample (Section 12.2's validation remark) – the 1-5% gap being the practical "price" of the $500\times$ scenario-count reduction, a trade-off operators accept given the alternative (an intractable daily re-planning problem) is not viable at all.

Proof sketch. This is the Lagrangian dual of the GMP: introduce multipliers $y_0 \in \mathbb{R}$ (for the normalization constraint $\int dP = 1$), $\mathbf{y}_1 \in \mathbb{R}^n$ (for the mean constraint), $\mathbf{Y}_2 \succeq 0$ (for the PSD covariance constraint, dual to a PSD cone is itself a PSD cone). The Lagrangian $\int \ell(\mathbf{x}, \boldsymbol{\xi}) dP - y_0(\int dP - 1) - \mathbf{y}_1^\top(\int \boldsymbol{\xi} dP - \boldsymbol{\mu}_0) - \operatorname{tr}(\mathbf{Y}_2(\int \boldsymbol{\xi}\boldsymbol{\xi}^\top dP - \mathbf{M}_0))$ must be bounded above over $P \succeq 0$, which forces the pointwise inequality constraint for every $\boldsymbol{\xi}$ in the support. ■

Example — One-dimensional worked dual

Let $\ell(\mathbf{x}, \xi) = (\xi - x)^2$ (squared loss, scalar), $\mu_0 = 0$, $M_0 = 1$ (unit second moment). The dual constraint is $y_0 + y_1\xi + y_2\xi^2 \geq (\xi - x)^2 = \xi^2 - 2x\xi + x^2$ for all $\xi \in \mathbb{R}$, i.e. $(y_2 - 1)\xi^2 + (y_1 + 2x)\xi + (y_0 - x^2) \geq 0$ for all ξ. This holds iff $y_2 \geq 1$ and the discriminant $(y_1 + 2x)^2 \leq 4(y_2 - 1)(y_0 - x^2)$ (for $y_2 > 1$; if $y_2 = 1$ then also $y_1 = -2x$ and $y_0 \geq x^2$). Minimizing $y_0 + y_1\mu_0 + y_2 M_0 = y_0 + y_2$ subject to these gives, at $y_2 = 1$, $y_1 = -2x$, $y_0 = x^2$, the dual value $x^2 + 1$. Direct computation confirms

$\sup_{P:\mathbb{E}_P[\xi]=0,\mathbb{E}_P[\xi^2]=1} \mathbb{E}_P[(\xi - x)^2] = \mathbb{E}_P[\xi^2] - 2x\mathbb{E}_P[\xi] + x^2 = 1 - 0 + x^2 = x^2 + 1$, matching the dual exactly – zero duality gap, as guaranteed by the theorem.

Multivariate generalization of the worked dual. For $\ell(\mathbf{x}, \boldsymbol{\xi}) = \|\boldsymbol{\xi} - \mathbf{x}\|_2^2$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\mathbf{M}_0 = \mathbf{I}$ ($d = \dim \boldsymbol{\xi}$), the dual value is $\|\mathbf{x}\|_2^2 + d$ by an analogous SOS argument. The pattern $\mathbb{E}_{P^*}[\|\boldsymbol{\xi} - \mathbf{x}\|^2] \leq \|\mathbf{x} - \boldsymbol{\mu}_0\|^2 + \text{tr}(\square_0)$ is the bias-variance decomposition in disguise – moment-DRO for quadratic loss reduces to classical bias-variance bounds.

Example — When the SOS relaxation is exact vs. a true relaxation

For ℓ quadratic and $\Xi = \mathbb{R}^d$ (unconstrained), the S-lemma-based SOS reformulation of "$\forall \boldsymbol{\xi} \in \mathbb{R}^d$" is *exact*. If instead Ξ is constrained (e.g. $\boldsymbol{\xi} \geq \mathbf{0}$), the pointwise constraint requires a *generalized* S-lemma which is generally only a *relaxation* for $d \geq 2$ – a gap analogous to the affine-decision-rule suboptimality of Section 9.5.

Computational note. The dual SDP of Theorem 10.3 has $O(d^2)$ variables regardless of N – a contrast to the $O(N)$-scenario extensive forms of Section 12.1–12.2. This N-independence is the key advantage of moment-based DRO for problems with few features d but large/streaming N, complementing Wasserstein's suitability for N-comparable, low-d problems.

Connections. The SDP dual derived here is the moment-based analogue of the Wasserstein dual of Section 10.2: in both cases, an infinite-dimensional worst-case-distribution problem collapses to a finite-dimensional dual over "shadow prices" for the ambiguity-set constraints (moments, or transport budget). The S-lemma / SOS step generalizes the support-function calculus of Section 9.1–9.3 from sets in $\mathbb{R}^n$ to sets of *measures*, and is the technical core that makes Part 10's results computationally implementable in standard conic solvers.

Theorem: complementary slackness in the GMP/SDP dual

At a primal-dual optimal pair $(P^*, (y_0^*, \mathbf{y}_1^*, \mathbf{Y}_2^*))$ for the GMP of Theorem 10.3 (assuming strong duality), the dual constraint $y_0^* + \mathbf{y}_1^{*\top}\boldsymbol{\xi} + \boldsymbol{\xi}^\top \mathbf{Y}_2^* \boldsymbol{\xi} \geq \ell(\mathbf{x}, \boldsymbol{\xi})$ holds with *equality* for P^*-almost-every $\boldsymbol{\xi}$ in the support of P^* – i.e. the worst-case distribution P^* is supported exactly on the set where the dual "certificate" polynomial touches $\ell(\mathbf{x}, \cdot)$.

Proof sketch. This is the measure-theoretic analogue of LP complementary slackness: if the dual constraint were strict ($>$) on a set of positive P^*-measure, then $\int [\text{dual poly} - \ell]\, dP^* > 0$, but by strong duality $\int [\text{dual poly}]\, dP^* = \text{dual objective} = \text{primal objective} = \int \ell\, dP^*$, a contradiction. ■

Example — Identifying the worst-case distribution's support

Recall the scalar example: $\ell(x, \xi) = (\xi - x)^2$, $\mu_0 = 0, M_0 = 1$, dual optimum $y_0^* = x^2, y_1^* = -2x, y_2^* = 1$, dual value $x^2 + 1$. The dual polynomial is $y_0^* + y_1^*\xi + y_2^*\xi^2 = x^2 - 2x\xi + \xi^2 = (\xi - x)^2 = \ell(x, \xi)$ – the dual polynomial *equals* ℓ *everywhere* (not just on a sub-support), so complementary slackness places *no restriction* on P^*'s support beyond the moment constraints themselves – consistent with the fact that *any* P with $\mathbb{E}_P[\xi] = 0, \mathbb{E}_P[\xi^2] = 1$ achieves $\mathbb{E}_P[(\xi - x)^2] = x^2 + 1$ exactly (as computed directly in Section 10.3's original worked example), i.e. the worst case is attained by *every* feasible P, not a unique one – a degenerate (but illustrative) case of complementary slackness where the dual certificate is tight identically, reflecting the quadratic loss's special compatibility with second-moment ambiguity.

Remarks and extensions. The GMP/SDP duality of this section underlies not only DRO but also the broader field of *polynomial optimization* (Lasserre's moment-SOS hierarchy): a polynomial optimization problem $\min_{\mathbf{x}} p(\mathbf{x})$ s.t. $g_j(\mathbf{x}) \geq 0$ can be relaxed to a hierarchy of SDPs of increasing size, each exact in the limit, by representing the (unknown) optimal $\mathbf{x}^*$ via its "moment sequence" $\{\mathbb{E}_{\delta_{\mathbf{x}^*}}[\mathbf{x}^{\boldsymbol{\alpha}}]\}_{\boldsymbol{\alpha}}$ and dualizing against sum-of-squares certificates – formally identical to the GMP duality of Theorem 10.3, but with the "distribution" P now a (degenerate) point mass $\delta_{\mathbf{x}^*}$ rather than a genuine ambiguity-set element. This connection means that advances in either field (polynomial optimization or moment-DRO) tend to transfer to the other, and the same SDP solvers (and the same numerical conditioning concerns as discussed in Section 9.3's remarks) apply to both.

A second remark: the GMP dual of Theorem 10.3 assumed the ambiguity set was defined by mean and covariance constraints. Higher-moment ambiguity sets (matching third or fourth moments, e.g. to capture skewness or kurtosis of $\boldsymbol{\xi}$) lead to dual problems involving degree-3 or degree-4 polynomial multipliers $y_0 + \mathbf{y}_1^\top \boldsymbol{\xi} + \dots$, and the corresponding S-lemma generalisation (the *Positivstellensatz*) requires higher levels of the Lasserre hierarchy – tractable in principle but with SDP sizes growing rapidly (as $O(d^{2k})$ for degree-$2k$ multipliers), illustrating a genuine (not merely conservatism-related) tractability frontier as ambiguity-set complexity increases.

Theorem: Lasserre hierarchy convergence for moment-DRO

Let $z^{(k)}$ be the optimal value of the degree-$2k$ SOS relaxation of the GMP dual. The sequence $z^{(k)}$ is non-increasing in k (tighter relaxations give smaller upper-bound certificates on the dual minimization) and converges to the true worst-case value $z^{(\infty)} = \sigma_{\mathcal{P}_{\text{mom}}}$; under Putinar's Archimedean condition, $z^{(k)} = z^{(\infty)}$ for some *finite* k whenever $\ell(\mathbf{x}, \cdot)$ has degree $\leq 2k$ and Ξ's defining polynomials have degree $\leq 2k$.

Proof sketch. Each relaxation level adds SOS-multiplier flexibility of higher degree, so higher-k dual feasible regions are *larger* (more certificates available), giving non-increasing optimal values $z^{(k)} \geq z^{(k+1)} \geq \dots \geq z^{(\infty)}$. Finite convergence under Putinar's condition is Lasserre's (2001) original finite-convergence theorem, transplanted to the GMP-dual setting. ■

Example — When $k = 1$ already suffices

For $\ell(x, \xi) = (\xi - x)^2$ (degree-2) and $\Xi = \mathbb{R}$ (no support constraints), the relevant SOS certificate is itself degree-2 – $k = 1$ suffices, $z^{(1)} = z^{(\infty)} = x^2 + 1$ exactly, with *no* relaxation gap, consistent with the "exact for unconstrained quadratic" remark of Section 10.3. For degree-4 ℓ (e.g. quartic loss, occasionally used for heavier outlier-penalization), $k = 1$ generally does *not* suffice, and $k = 2$ (degree-4 SOS, $O(d^2)$-sized PSD matrices vs. $O(d)$ for $k = 1$) is needed – the hierarchy's first non-trivial step beyond the quadratic case used throughout this volume's examples.

Historical notes. The generalized moment problem and its SDP dual (Theorem 10.3) sit at the intersection of two major 20th/21st-century mathematical programs: the classical *moment problem* in functional analysis (Hamburger, Stieltjes, Hausdorff, early 20th century, asking when a sequence of numbers can be the moments of some measure), and the more recent *polynomial optimization / sum-of-squares* program (Lasserre, 2001; Parrilo, 2003), which recognised that the moment problem's dual – certificates of non-negativity for polynomials – could be computed via SDP using the Positivstellensatz (a real-algebraic-geometry result with roots in Hilbert's 17th problem, 1900). The application of this machinery specifically to distributionally robust optimization (rather than to polynomial optimization more broadly)

is comparatively recent (Delage & Ye, 2010, is an early and influential reference), making Section 10.3 perhaps the section of this volume with the deepest mathematical lineage relative to its optimization-application's youth.

Theorem: zero duality gap for moment-DRO under a Slater-type condition

For the generalized moment problem dual of Section 10.3 with ambiguity set $\mathcal{P}_{\text{mom}} = \{P \succeq 0 : \int dP = 1, \int g_j(\xi) dP = m_j,\ j = 1, \dots, J\}$ (finitely many generalized moment constraints), if there exists a probability measure P_0 strictly satisfying all J moment constraints in their interior (a Slater point) and the loss $\ell(\mathbf{x}, \cdot)$ is bounded and measurable, then the primal supremum and the SDP/LP dual infimum coincide exactly – no duality gap, regardless of J or the dimension of Ξ.

Proof sketch. This is an instance of the general infinite-dimensional conic duality theorem (e.g. via the moment-SOS hierarchy's strong duality results, Lasserre 2009), specialized to the case of finitely many linear moment constraints on a measure: Slater's condition for the (infinite-dimensional, but conic) primal moment problem ensures the dual achieves the same optimal value, by the same mechanism as finite-dimensional Slater-based strong conic duality (Theorem 9.3's addendum), here applied in the measure-space setting. ■

Example — Verifying Slater's condition in the worked moment example

Recall the worked $\mathcal{P}_{\text{mom}} = \{P : \mathbb{E}_P[\xi] = \mu_0, \text{Cov}_P(\xi) \preceq \Sigma_0\}$ example of Section 10.3's main text (with $\mu_0 = \mathbf{0}$, $\Sigma_0 = \mathbf{I}$, $d = \dim \xi$). The standard Gaussian $\mathcal{N}(\mathbf{0}, 0.5\mathbf{I})$ satisfies $\mathbb{E}[\xi] = \mathbf{0} = \mu_0$ exactly and $\text{Cov} = 0.5\mathbf{I} \prec \mathbf{I} = \Sigma_0$ *strictly* (in the PSD order) – a valid Slater point. By the theorem, the dual value computed in Section 10.3's worked example ($\|\mathbf{x}\|_2^2 + d$) is therefore not merely an upper bound but the *exact* primal supremum, with no gap – confirming numerically what the SOS-based derivation already asserted constructively, but now via a general existence argument that extends to ambiguity sets too complex for an explicit SOS certificate to be written down by hand.

When Slater's condition fails. If the moment constraints are exactly binding with no interior point (e.g. an ambiguity set defined by an *equality* second-moment constraint $\text{Cov}_P(\xi) = \Sigma_0$ rather than $\preceq$, with Σ_0 singular), Slater's condition can fail and a duality gap, while rare in practice, becomes a theoretical possibility; in such degenerate cases the primal supremum may exceed the dual SDP value, and the practical recommendation – standard in the SDP-duality literature – is to perturb the ambiguity set slightly (e.g. relax the equality to $\Sigma_0 + \delta\mathbf{I}$ for small $\delta > 0$) to restore strict interior feasibility before solving, accepting a small, controllable conservatism in exchange for a duality-gap-free, hence tighter and more interpretable, dual reformulation.

Historical notes (addendum). The application of Slater-type conditions to guarantee zero duality gap specifically in the moment-DRO setting traces to the generalized moment problem literature originating with Isii (1960, already cited in Section 12.3's historical notes) and developed into modern SDP language by Lasserre's moment-SOS hierarchy programme (2001 onward); the explicit Slater-point verification technique illustrated in the worked example above is a standard but underused diagnostic that practitioners can apply to any specific moment-DRO instance before trusting a numerically computed dual value as exact.

10.4 Data-Driven Ambiguity Sets and Sample Complexity

Overview

For DRO to be statistically meaningful, the ambiguity set radius ϵ must be chosen so that $\mathcal{P}$ contains the true (unknown) distribution P^* with high probability, given only N i.i.d. samples. This section derives such finite-sample guarantees via concentration inequalities, connecting the geometric size of $\mathcal{P}$ to the sample size N.

Example — Case study: structural reliability with chance-constrained load factors

A building code requires Pr[load $\leq$ capacity] ≥ 0.999 (a $\beta = 0.001$ chance constraint, reflecting a roughly "1-in-1000-year" acceptable failure probability per relevant load combination). If load is modeled as Gaussian with mean $\bar{L}$ and std. dev. σ_L, Theorem 12.3 gives the required capacity $C \geq \bar{L} + \Phi^{-1}(0.999)\sigma_L \approx \bar{L} + 3.09\sigma_L$ – the "3.09-sigma" safety margin one sometimes sees cited in structural codes. If instead a Chebyshev (distribution-free) bound is used (e.g. because load distributions for extreme events are known to be heavy-tailed, non-Gaussian), the required margin becomes $\bar{L} + \sqrt{(1-0.001)/0.001}\,\sigma_L \approx \bar{L} + 31.6\sigma_L$ – an order of magnitude larger. This $10\times$ gap between Gaussian and Chebyshev margins for small β is precisely the numerical pattern illustrated abstractly in the main text, here with direct consequences for material costs (a 31.6σ margin would be economically infeasible for most structures) – in practice, structural codes instead use *empirically-fitted* extreme-value distributions (Volume I, Extreme Values & Tail Risk) for L's tail, avoiding both the Gaussian's potential under-protection and the Chebyshev bound's crippling over-conservatism.

Concentration of the empirical distribution. The key quantity is $W_p(\hat{P}_N, P^*)$, the Wasserstein distance between the empirical distribution of N samples from P^* and P^* itself.

Theorem: Wasserstein concentration (Fournier & Guillin, 2015, simplified)

If $\boldsymbol{\xi} \sim P^*$ has support in a bounded set of diameter D in $\mathbb{R}^d$, then for $\beta \in (0,1)$, with probability at least $1-\beta$ over the draw of N i.i.d. samples,

$$W_1(\hat{P}_N, P^*) \leq \epsilon_N(\beta) \triangleq C\,D\left(\frac{1}{N}\log\frac{1}{\beta}\right)^{1/\max(d,2)}$$

for a universal constant $C > 0$ depending only on d.

Example — Case study: a hydroelectric reservoir management problem

A hydroelectric utility manages reservoir water levels over $T = 12$ monthly stages, with uncertain monthly inflows ξ_t (rainfall-dependent) and a value-to-go function $V_t(\text{reservoir level})$ capturing the expected future value of water in storage. SDDP is the standard industrial tool for such problems (used by grid operators in Brazil, Norway, and elsewhere for exactly this application) precisely because the reservoir-level state is low-dimensional (often just 1-5 reservoirs $\Rightarrow$ state $\in \mathbb{R}^{1\text{-}5}$) while the inflow uncertainty, over $T = 12$ or more stages with multiple inflow scenarios per stage, would give an utterly intractable extensive form ($O(b^{11})$ for $b \geq 3$ branches/stage, as in the Section 12.4 tree-size example). SDDP's per-iteration cost depends on the (small) state dimension, not on b^{T-1}, making the problem tractable – this real-world application is

the historical motivation for SDDP's development (Pereira & Pinto, 1991) and remains its largest-scale deployment.

Sample complexity and the curse of dimensionality. Setting $\mathcal{P} = \{P : W_1(P, \hat{P}_N) \le \epsilon_N(\beta)\}$ guarantees, by the triangle inequality $W_1(P^*, \hat{P}_N) \le \epsilon_N(\beta)$ (with probability $\ge 1-\beta$), that $P^* \in \mathcal{P}$, hence

$$\mathbb{E}_{P^*}[\ell(\mathbf{x}^*, \boldsymbol{\xi})] \le \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\mathbf{x}^*, \boldsymbol{\xi})]$$

for the WDRO-optimal $\mathbf{x}^*$: the WDRO objective value is, with probability $\ge 1-\beta$, an upper bound (a guarantee) on the true expected loss. However, $\epsilon_N(\beta) \to 0$ only at rate $N^{-1/\max(d,2)}$ – for $d \ge 3$ this is slower than the parametric $N^{-1/2}$ rate, the manifestation of the curse of dimensionality in this finite-sample guarantee.

Example — Sample size for a target guarantee

Suppose $d = 10$, $D = 1$, $\beta = 0.05$ ($\log(1/\beta) \approx 3$), and we want $\epsilon_N(\beta) \le 0.1$. Then $\epsilon_N(\beta) = C(3/N)^{1/10} \le 0.1 \Rightarrow (3/N) \le (0.1/C)^{10}$. For $C \approx 1$ (order of magnitude), $(3/N) \le 10^{-10}$, i.e. $N \ge 3 \times 10^{10}$ – an astronomically large sample size, illustrating that for high-dimensional $\boldsymbol{\xi}$, the *worst-case-over-$\mathcal{P}$* guarantee with small ϵ is statistically vacuous unless additional structure (e.g. low-dimensional sufficient statistics, as in the moment-based sets of Section 10.1, which concentrate at the parametric rate $N^{-1/2}$ regardless of d) is exploited.

Why moment-based sets avoid the curse. The dimension-independence of moment-based DRO extends to its statistical guarantee: $\hat{\boldsymbol{\mu}}_N \to \boldsymbol{\mu}^*$ and $\hat{\square}_N \to \square^*$ at rate $O_p(N^{-1/2})$ *regardless of* d. Setting the moment-ambiguity radius at this scale gives a valid finite-sample guarantee without the $N^{-1/d}$ degradation of Theorem 10.4, at the cost of only protecting against distributions sharing the assumed moments.

Example — Numerical contrast: sample sizes for $d = 10$, target $\epsilon = 0.1$

For $d = 10$, $\beta = 0.05$, target Wasserstein radius 0.1 required $N \gtrsim 3 \times 10^{10}$. For a moment-based confidence ellipsoid on $(\boldsymbol{\mu}, \square) \in \mathbb{R}^{10} \times \mathbb{R}^{10\times 10}_{\text{sym}}$ (≈ 65 free parameters), radius $\propto \sqrt{65/N} \le 0.1$ requires $N \gtrsim 6500$ – seven orders of magnitude smaller. Always match the ambiguity-set family to the available sample size.

Adaptive radius selection. In practice, ϵ is often chosen by cross-validation: solve DRO for a grid of ϵ, evaluate held-out performance, select ϵ^*. This sidesteps conservative theoretical bounds while retaining DRO's structural benefit – $\mathbf{x}^*(\epsilon^*)$ is still the solution of a regularized problem (Section 10.5) with a data-calibrated strength, analogous to cross-validated λ in ridge/LASSO.

Data-driven ϕ-divergence radii. For KL-divergence ambiguity sets, a similar finite-sample story holds via Sanov's theorem / the method of types: $\Pr[D_{\text{KL}}(\hat{P}_N \,\|\, P^*) > \epsilon] \le (N+1)^{|\Xi|} e^{-N\epsilon}$ for finite support Ξ, giving $\epsilon_N(\beta) = O(\log(1/\beta)/N)$ – the parametric N^{-1} rate (faster than even $N^{-1/2}$ for the divergence itself, though the implied radius in distribution space behaves differently). This dimension-independence (depending on $|\Xi|$ rather than continuous d) is a key practical advantage of ϕ-divergence over Wasserstein ambiguity sets in discrete or discretized settings, and motivates the choice of ambiguity-set family based on the structure of $\boldsymbol{\xi}$ – a decision that, together with the regularization-equivalence of Section 10.5, completes the practical toolkit for applying DRO.

Theorem: union bound for simultaneous coverage across multiple candidate solutions

If a decision-maker evaluates K candidate solutions $\mathbf{x}_1, \dots, \mathbf{x}_K$ (e.g. from a grid search over ϵ, or different model architectures) and wants *all* of them to satisfy $\mathbb{E}_{P^*}[\ell(\mathbf{x}_k, \boldsymbol{\xi})] \leq \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\mathbf{x}_k, \boldsymbol{\xi})]$ simultaneously with overall confidence $\geq 1 - \beta$, it suffices to use $\mathcal{P} = \{P : W_1(P, \hat{P}_N) \leq \epsilon_N(\beta/K)\}$ – a Bonferroni-corrected radius, the WDRO analogue of the Bonferroni joint-chance-constraint correction of Section 12.3.

Proof. By Theorem 10.4 applied with confidence level β/K, $\Pr[W_1(\hat{P}_N, P^*) > \epsilon_N(\beta/K)] \leq \beta/K$. By the union bound, $\Pr[\exists k : \text{coverage fails for } \mathbf{x}_k] \leq K \times (\beta/K) = \beta$ (since coverage for *every* $\mathbf{x}_k$ depends only on the *single* event $W_1(\hat{P}_N, P^*) \leq \epsilon_N(\beta/K)$, which either holds for all k simultaneously or fails for all k simultaneously – so the union bound here is actually not even needed in its full generality, but the β/K radius is the conservative choice that would be needed if coverage depended on k-specific random quantities). ■

Example — Multiple-testing-style correction for a hyperparameter grid

A grid search over $K = 20$ values of ϵ is performed, and the analyst wants 95% *overall* confidence ($\beta = 0.05$) that the selected ϵ^*'s WDRO bound is valid. Using $\beta/K = 0.0025$ in place of $\beta = 0.05$ in $\epsilon_N(\beta) = CD(\frac{1}{N}\log\frac{1}{\beta})^{1/\max(d,2)}$: for $d = 2$, $\log(1/0.05) \approx 3$ vs. $\log(1/0.0025) \approx 6$ – a factor-of-2 increase inside the (square-root, for $d = 2$) radius, i.e. $\epsilon_N(\beta/K)/\epsilon_N(\beta) = \sqrt{6/3} = \sqrt{2} \approx 1.41$. A 41% larger radius is the (modest, for $d = 2$) cost of correcting for the $K = 20$-fold search – for larger d (slower-decaying log-dependence under the $1/\max(d, 2)$ exponent... actually the exponent affects how the radius scales with N, while the $\log(1/\beta)$ factor enters linearly inside the parenthesis regardless of d, so the *relative* correction factor $(\log(1/\beta_2)/\log(1/\beta_1))^{1/\max(d,2)}$ *shrinks* toward 1 as d grows) – the multiple-testing correction is proportionally *smaller* in relative terms for high-dimensional problems, a small silver lining to the curse of dimensionality.

Remarks and extensions. The Wasserstein concentration bound of Theorem 10.4 is known to be *tight* (up to constants) for $\boldsymbol{\xi}$ uniformly distributed on $[0, 1]^d$ – the empirical measure of N uniform samples genuinely is at W_1-distance $\Theta(N^{-1/d})$ from the uniform distribution for $d \geq 3$, so the curse of dimensionality identified here is not an artefact of a loose bound but a genuine statistical fact about empirical measures in high dimensions. This is closely related to the well-known difficulty of high-dimensional density estimation, and explains why purely nonparametric (distribution-free, Wasserstein-based) approaches to DRO are, in practice, confined to low-dimensional $\boldsymbol{\xi}$ (often $d \leq 5$ or so) unless combined with dimension-reduction (e.g. applying WDRO not to raw $\boldsymbol{\xi} \in \mathbb{R}^d$ but to a low-dimensional sufficient statistic or learned embedding $T(\boldsymbol{\xi}) \in \mathbb{R}^{d'}$, $d' \ll d$).

A practically important remark: the sample-complexity results of this section are *worst-case over all P^* with bounded support*. For specific, well-behaved P^* (e.g. sub-Gaussian with rapidly decaying tails, or supported on a low-dimensional manifold embedded in $\mathbb{R}^d$), the *effective* dimension governing the $W_1(\hat{P}_N, P^*)$ rate can be much smaller than the ambient d – a phenomenon sometimes called the "intrinsic dimension" of the data, and an active research area connecting statistical DRO theory to manifold-learning ideas from the companion Neural Networks series.

Theorem: sequential (online) ambiguity-radius shrinkage

For sequential data arrival $\hat{P}_N \to \hat{P}_{N+1}$, the Wasserstein radius ratio satisfies $\epsilon_{N+1}(\beta)/\epsilon_N(\beta) = (N/(N+1))^{1/\max(d,2)} \approx 1 - \frac{1}{N\max(d,2)}$ for large N – a small, predictable per-sample shrinkage, enabling a "warm-start" DRO-update strategy rather than full re-solves.

Proof. Direct from $\epsilon_N(\beta) \propto N^{-1/\max(d,2)}$: $(N/(N+1))^a = (1 - \frac{1}{N+1})^a \approx 1 - \frac{a}{N}$ for $a = 1/\max(d,2)$ and large N. ■

Example — Online radius updates for a streaming fraud model

For $d = 5$ (dimension-reduced embedding, Section 10.4's case study), $\max(d,2) = 5$, so $\epsilon_{N+1}/\epsilon_N \approx 1 - \frac{1}{5N}$. At $N = 10000$: ratio ≈ 0.99998 – negligible per-sample shrinkage, meaning the DRO solution can be left unchanged for long stretches. Over $\Delta N = 1000$ new samples, cumulative shrinkage $\approx (0.99998)^{1000} \approx 0.98$, a 2% radius reduction – perhaps the threshold at which a practitioner triggers a re-solve. This "shrinkage budget" reasoning gives a principled, computation-aware re-solve schedule for streaming DRO, an operational concern beyond the static sample-complexity theory of Section 10.4.

Historical notes. Concentration-of-measure results for empirical distributions (the Fournier–Guillin theorem underlying Theorem 10.4) belong to the broader 20th-century theory of empirical processes (Vapnik & Chervonenkis, 1970s; Dudley, 1970s-80s), the same body of theory underlying VC-dimension-based generalization bounds in statistical learning theory – so the "curse of dimensionality" identified in Section 10.4 for WDRO sample complexity is, mathematically, the *same* curse that affects nonparametric density estimation and certain generalization bounds discussed in the companion Neural Networks series' learning-theory chapters. Fournier and Guillin's specific 2015 result sharpened earlier (1990s-2000s) bounds to the (now standard) rates quoted in this section, and its application to calibrating WDRO ambiguity radii (rather than to its original empirical-process-theory motivations) is a direct illustration of mathematical results finding new application decades after their original development – a recurring pattern across this volume's historical notes.

Theorem: bootstrap calibration of the Wasserstein radius as an alternative to concentration bounds

Let $\hat{\epsilon}_N(\beta)$ be the $(1-\beta)$-quantile of the bootstrap distribution of $W_1(\hat{P}_N, \hat{P}_N^{(b)})$, where $\hat{P}_N^{(b)}$ denotes a resample of size N drawn with replacement from $\hat{P}_N$ (the standard nonparametric bootstrap), repeated over B bootstrap replications. Under standard bootstrap consistency conditions (Section 10.4's main-text concentration-bound regularity plus a mild smoothness condition on the loss), $\hat{\epsilon}_N(\beta) \to \epsilon_N^\star(\beta)$ in probability as $B \to \infty$, where $\epsilon_N^\star(\beta)$ is the (generally unknown, finite-sample) exact radius achieving coverage $1-\beta$ – the bootstrap radius is an asymptotically valid, often substantially tighter, alternative to the worst-case concentration-inequality-based radius of Theorem 10.4's main-text result.

Proof sketch. This is a direct application of the standard nonparametric-bootstrap consistency theory (Efron 1979; Bickel–Freedman 1981) to the specific functional $\hat{P}_N \mapsto W_1(\hat{P}_N, P^\star)$, treating the bootstrap's resampling distribution as a proxy for the true sampling distribution of this functional around $\hat{P}_N$ – valid under the same smoothness/moment conditions required for bootstrap consistency of other $N^{-1/2}$-rate functionals, here applied to the em-

pirical Wasserstein distance rather than a simple mean or variance. ■

Example — Why the bootstrap radius is typically smaller

The concentration-bound radius of Theorem 10.4 (main text) is derived from a worst-case, distribution-free argument valid for *any* distribution satisfying the stated moment/support conditions, and is consequently conservative for any specific dataset whose empirical higher moments (skewness, kurtosis) happen to be more favourable than the worst case the bound accounts for. The bootstrap radius $\hat{\epsilon}_N(\beta)$, by contrast, is calibrated directly from the *observed* data's empirical distribution, automatically adapting to favourable features of the specific sample at hand. In a numerical comparison on simulated sub-Gaussian data with $N = 200$, $d = 5$, $\beta = 0.1$: the concentration-bound radius (following Theorem 10.4's formula) might give $\epsilon_N \approx 0.85$, while a $B = 2000$-replication bootstrap calibration on the same data might give $\hat{\epsilon}_N(0.1) \approx 0.52$ – a substantial (here, nearly 40%) reduction in required ambiguity-set radius, translating directly into a less conservative (lower expected cost) DRO solution for the same nominal coverage target, at the price of losing the concentration bound's distribution-free worst-case guarantee in favour of an asymptotic, sample-dependent one.

A caution on small-sample bootstrap reliability. The bootstrap radius's asymptotic justification means its finite-sample reliability for small N (say, $N < 50$) should not be taken for granted, particularly in higher dimensions d where the empirical distribution $\hat{P}_N$ may itself be a poor proxy for $P^\star$'s tail behaviour purely due to limited sample coverage of the support – exactly the curse-of-dimensionality concern already raised for the Wasserstein sample-complexity rate $O(N^{-1/d})$ in this section's main text. A standard practical compromise is to use the bootstrap radius as a *point estimate* alongside the concentration bound as a conservative *upper envelope*, selecting a radius between the two (or simply the concentration bound when the two disagree substantially) as a pragmatic hedge against bootstrap under-coverage in finite, especially high-dimensional, samples.

Historical notes (addendum). The use of bootstrap resampling to calibrate DRO ambiguity-set radii directly, rather than relying solely on distribution-free concentration inequalities, reflects a broader and recurring tension in statistics between distribution-free worst-case guarantees (this volume's Part 9-10 emphasis) and data-adaptive, asymptotically-justified procedures (the classical statistical bootstrap, Efron 1979) – a tension also visible in the bias-variance radius-selection discussion of this section's earlier addendum, and one without a universally agreed resolution: the choice between the two calibration philosophies remains, in practice, a judgment call informed by sample size, dimension, and the consequences of potential ambiguity-set under-coverage in the application at hand.

Theorem: bias-variance decomposition of the Wasserstein-DRO out-of-sample gap

Let $z^\star(\epsilon) = \min_{\mathbf{x}} \sup_{Q: W_1(Q, \hat{P}_N) \le \epsilon} \mathbb{E}_Q[\ell(\mathbf{x}, \xi)]$ be the Wasserstein-DRO optimal value at radius ϵ, and $z^{\text{true}} = \min_{\mathbf{x}} \mathbb{E}_{P^\star}[\ell(\mathbf{x}, \xi)]$ the (unknown) true-distribution optimum. The out-of-sample gap decomposes as

$$z^\star(\epsilon) - z^{\text{true}} = \underbrace{\left(z^\star(\epsilon) - \mathbb{E}_{\hat{P}_N}[\ell(\mathbf{x}^\star_\epsilon, \xi)]\right)}_{\text{robustness margin } (\ge 0, \text{ increasing in } \epsilon)} + \underbrace{\left(\mathbb{E}_{\hat{P}_N}[\ell(\mathbf{x}^\star_\epsilon, \xi)] - z^{\text{true}}\right)}_{\text{estimation bias (sign varies, magnitude} \sim O_p(N^{-1/2}), \text{ roughly decreasing influence a}}$$

Proof sketch. Direct algebraic decomposition (add and subtract $\mathbb{E}_{\hat{P}_N}[\ell(\mathbf{x}^\star_\epsilon, \xi)]$); the first term is non-negative by definition of the sup, and increasing in ϵ by monotonicity of the ambiguity

ball in its radius; the second term's stochastic order follows from the standard empirical-process convergence rate $\hat{P}_N \to P^\star$ at rate $N^{-1/2}$ (in, e.g., bounded-Lipschitz metric) for fixed, N-independent $\mathbf{x}^\star_\epsilon$, with the qualitative "roughly decreasing influence" claim reflecting that larger ϵ values make $\mathbf{x}^\star_\epsilon$ increasingly conservative and less sensitive to the specific sample realization (though this is an informal, not a uniform, statement). ■

Example — Locating the bias-variance-optimal radius numerically

Following Section 10.4's main-text setup ($d = 10$, target out-of-sample reliability $\beta = 0.05$), suppose simulation reveals the robustness margin term grows roughly as 0.5ϵ while the estimation-bias term's typical magnitude decays roughly as $2/\sqrt{N}$ for the sample size N at hand (a stylized, illustrative rate). Setting these two terms equal, $0.5\epsilon^\star \approx 2/\sqrt{N}$, gives $\epsilon^\star \approx 4/\sqrt{N}$ as a heuristic "balanced" radius – smaller than this trades excess estimation bias for insufficient robustness margin, while larger than this over-protects against a margin of safety not justified by the data's actual estimation uncertainty at the given N. This heuristic balance point is distinct from (and typically smaller than) the formal finite-sample-guarantee radius of Theorem 10.4 in the main text, which is calibrated for a worst-case coverage guarantee rather than an expected-gap minimization – the two criteria (worst-case coverage vs. expected-gap minimization) generally recommend *different* radii, a distinction analogous to the VaR-vs-CVaR conservatism gradient noted in Section 12.3.

Practical guidance. In settings where the decision-maker cares about *average* (rather than worst-case) performance across repeated deployments of the same robust model (e.g. a recurring monthly planning problem, not a one-off high-stakes decision), the bias-variance-balanced radius is the more relevant target; in settings with a genuine one-shot, high-consequence decision, the worst-case-coverage radius of Theorem 10.4 remains the appropriate (more conservative) choice – the two should not be conflated, and a practitioner reporting a single "optimal ϵ" should specify which criterion was used.

A caveat on the decomposition's informality. It is worth being explicit that the decomposition above, while algebraically exact, packages a genuinely delicate statistical question (how does $\mathbf{x}^\star_\epsilon$'s sensitivity to the sample actually vary with ϵ?) into a single qualitative phrase ("roughly decreasing influence"). A fully rigorous version would need to track how the *set* of near-optimal $\mathbf{x}$ shrinks or grows with ϵ, not just the point estimate $\mathbf{x}^\star_\epsilon$ – a subtlety this section's heuristic balance point deliberately sets aside in favour of practical usability, but one a reader pursuing finite-sample-exact radius selection should not lose sight of.

10.5 Connections to Regularization and Adversarial Training

Overview

This section makes explicit the connection, foreshadowed in Section 10.2, between distributional robustness and two staples of modern statistical learning: norm-based regularization (ridge, LASSO, group-lasso) and adversarial training. The unifying lens is that all three can be written as $\min_{\mathbf{x}} \sup_{\delta \in \mathcal{B}} \mathbb{E}[\ell(\mathbf{x}, \boldsymbol{\xi} + \delta)]$ for an appropriate perturbation set $\mathcal{B}$.

Example — Case study: classifying a new risk-management problem using the unifying table

A logistics company must decide how to model uncertain delivery times for a routing optimization. *Step 1*: is the uncertainty about a *parameter* (e.g. travel time on each road segment, support roughly known from historical extremes) or a *distribution* (e.g. the distribution of *daily* total travel time, which the company has rich historical data for but worries may shift with new traffic patterns)? Here, both apply at different levels: segment-level travel times are naturally *parameter*-uncertain (Part 9/left column), while daily aggregate patterns are *distribution*-uncertain (Part 10/bottom row). *Step 2*: what guarantee is needed – almost-sure (e.g. "never miss a hard delivery deadline") or probabilistic (e.g. "95% on-time rate")? Hard deadlines → top row (RO/DRO); soft SLAs → bottom row (CCP/SP). A realistic routing model might therefore combine: RO (Part 9) for hard per-segment deadlines using box/budgeted uncertainty (Sections 9.1, 9.4), *and* a CCP (Section 12.3) for the soft daily on-time-rate SLA using a distribution fitted to historical aggregate data – exactly the additively-combined RO+SP structure of this section's main worked example, now derived "top-down" from the business requirements via the unifying table, rather than "bottom-up" from a pre-specified mathematical structure. This top-down classification procedure is, in essence, the practical takeaway of the entire volume.

Adversarial training as pointwise WDRO. Adversarial training solves $\min_{\mathbf{x}} \frac{1}{N}\sum_i \max_{\|\delta_i\|\le\epsilon} \ell(\mathbf{x}, \hat{\boldsymbol{\xi}}_i + \delta_i)$, i.e. a *pointwise* (per-sample) worst-case perturbation, rather than a distributional one. This is the ∞-Wasserstein ($p=\infty$) limit of WDRO: $W_\infty(P, \hat{P}_N) \le \epsilon$ restricted to couplings supported on $\{(\hat{\boldsymbol{\xi}}_i, \hat{\boldsymbol{\xi}}_i + \delta_i) : \|\delta_i\| \le \epsilon\}$ is exactly the adversarial-training feasible set, with the outer sup over such P reducing to the per-sample inner maximizations.

Theorem: Regularization as the $\epsilon \to 0$ expansion of adversarial risk

For $\ell(\mathbf{x}, \boldsymbol{\xi})$ differentiable in $\boldsymbol{\xi}$ and $\mathcal{B} = \{\delta : \|\delta\|_q \le \epsilon\}$,

$$\max_{\|\delta\|_q \le \epsilon} \ell(\mathbf{x}, \boldsymbol{\xi} + \delta) = \ell(\mathbf{x}, \boldsymbol{\xi}) + \epsilon \left\|\nabla_{\boldsymbol{\xi}} \ell(\mathbf{x}, \boldsymbol{\xi})\right\|_{q^*} + O(\epsilon^2),$$

where q^* is the Hölder conjugate of q ($1/q + 1/q^* = 1$). Hence the adversarial risk equals the nominal risk plus a first-order penalty on the *dual-norm of the input gradient*, a generalized gradient-regularization term.

Proof sketch. A first-order Taylor expansion gives $\ell(\mathbf{x}, \boldsymbol{\xi} + \delta) \approx \ell(\mathbf{x}, \boldsymbol{\xi}) + \nabla_{\boldsymbol{\xi}} \ell(\mathbf{x}, \boldsymbol{\xi})^\top \delta$. Maximizing the linear term over $\|\delta\|_q \le \epsilon$ gives, by the Hölder/Cauchy–Schwarz duality identity $\sup_{\|\delta\|_q \le \epsilon} \mathbf{g}^\top \delta = \epsilon \|\mathbf{g}\|_{q^*}$ (the same identity underlying Eq. (9.1.ell) for $q = q^* = 2$), the stated result. ■

Example — From adversarial training to weight decay in linear models

For $\ell(\mathbf{x}, (\mathbf{u}, v)) = \frac{1}{2}(v - \mathbf{x}^\top \mathbf{u})^2$, $\nabla_{\mathbf{u}} \ell = -(v - \mathbf{x}^\top \mathbf{u})\mathbf{x}$, so $\|\nabla_{\mathbf{u}} \ell\|_{q^*} = |v - \mathbf{x}^\top \mathbf{u}| \cdot \|\mathbf{x}\|_{q^*}$. The adversarial (input-perturbation) risk to first order is $\frac{1}{2}(v - \mathbf{x}^\top \mathbf{u})^2 + \epsilon |v - \mathbf{x}^\top \mathbf{u}| \, \|\mathbf{x}\|_{q^*}$. For small residuals this is dominated by the quadratic term plus a term linear in $\|\mathbf{x}\|_{q^*}$ – for $q^* = 2$, recognisable as an approximate ridge penalty on $\mathbf{x}$ (matching the exact WDRO-regression result of Section 10.2's example), and for $q^* = 1$ as an approximate LASSO penalty. Numerically, with $\mathbf{x} = (1, -2)^\top$, $q^* = 2$: $\|\mathbf{x}\|_2 = \sqrt{5} \approx 2.236$, so a

unit increase in ϵ adds $\approx 2.236|v - \mathbf{x}^\top \mathbf{u}|$ to the loss – a perturbation-budget-dependent, residual-dependent penalty distinct from the constant $\lambda \|\mathbf{x}\|_2^2$ of standard ridge.

Higher-order terms and certified robustness. The first-order expansion of Theorem 10.5 is exact only as $\epsilon \to 0$. For finite ϵ, certified bounds give $\max_{\|\delta\| \le \epsilon} \ell(\mathbf{x}, \boldsymbol{\xi} + \delta) \le \ell(\mathbf{x}, \boldsymbol{\xi}) + \epsilon \|\nabla \ell\|_{q^*} + \frac{1}{2}\epsilon^2 \sup \|\nabla^2 \ell\|_{\text{op}} + \ldots$ – connecting adversarial robustness to loss-landscape curvature, a major theme of the companion Mathematics of Neural Networks series.

Example — Gradient norm as an empirical robustness diagnostic

Two models with identical training loss 0.05 but $\|\nabla_{\boldsymbol{\xi}} \ell\|_2 = 0.1$ (M_1) vs. 2.0 (M_2). With $\epsilon = 0.05$: M_1's adversarial risk ≈ 0.055 (10% increase); M_2's ≈ 0.15 (200% increase). Identical clean accuracy, dramatically different robustness – a cheap single-backward-pass diagnostic directly justified by the DRO-regularization equivalence.

Beyond ℓ_p perturbations. Real-world distribution shift is rarely well-modelled by a small ℓ_p ball. WDRO generalizes by choosing $d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ to be a semantic/perceptual distance – Theorem 10.2 holds for *any* metric d, so the ambiguity-set/regularizer dictionary extends to semantic robustness with $\|\cdot\|_{q^*}$ replaced by the dual norm in the relevant feature space – the conceptual culmination of Part 10.

Synthesis of Part 10. The chain of results across Section 10.1–10.5 establishes a single coherent picture: an ambiguity set $\mathcal{P}$ around the empirical/training distribution (Section 10.1) admits, via conic or transport duality (Sections 10.2–10.3), a tractable dual that is exactly an empirical-risk-plus-penalty objective; the penalty's form (norm of $\mathbf{x}$, norm of input gradients, or moment-based SDP terms) depends on the geometry of $\mathcal{P}$, and the *size* of $\mathcal{P}$ that yields valid finite-sample guarantees (Section 10.4) determines the penalty's strength. This dictionary – ambiguity-set geometry $\leftrightarrow$ regularizer – is the central conceptual takeaway connecting robust and distributionally robust optimization (Parts 9–10) to the decision-theoretic frameworks of Part 11 and the explicitly multi-period stochastic models of Part 12.

Part 11

Decision Theory under Deep Uncertainty

Beyond expected utility: formal frameworks for decisions when probabilities themselves are unknown or contested.

Theorem: equivalence class of ambiguity sets yielding the same regularizer

Two ambiguity sets $\mathcal{P}_1, \mathcal{P}_2$ (possibly of different "types" – e.g. one Wasserstein, one ϕ-divergence) yield the *same* first-order regularization penalty $\epsilon \|\nabla \ell\|_{q^*}$ if and only if their respective "radius-to-penalty" maps agree to first order at $\epsilon = 0$ – i.e. $\lim_{\epsilon \to 0} \frac{1}{\epsilon}\left[\sup_{P \in \mathcal{P}_i(\epsilon)} \mathbb{E}_P[\ell] - \mathbb{E}_{\hat{P}_N}[\ell]\right]$ is the same for $i = 1, 2$. This limit, for Wasserstein balls with cost d^q, equals $\|\nabla \ell\|_{q^*}$ (Theorem 10.5); for ϕ-divergence balls it equals $\sqrt{2\phi''(1)\mathrm{Var}_{\hat{P}_N}(\ell)}$ (a different functional form, generally *not* equal to $\|\nabla \ell\|_{q^*}$ except for special (ϕ, ℓ) pairs).

Proof sketch. Both limits are, by construction, the directional derivative of the "worst-case expectation" functional at $\epsilon = 0$ in the direction of increasing ambiguity-set radius. For Wasserstein balls this derivative is computed via the Kantorovich-duality argument of Theorem 10.2's proof, linearized; for ϕ-divergence balls it is computed via a second-order Taylor

expansion of ϕ around 1 (since $D_\phi(P\|\hat{P}_N) = 0$ at $P = \hat{P}_N$, and ϕ is minimized at 1 with $\phi(1) = 0, \phi'(1) = 0$ for valid ϕ-divergences, so the leading term is the *quadratic* $\phi''(1)$ term) – the two derivations involve genuinely different convex-analytic machinery and need not agree. ■

Example — When Wasserstein and χ^2 regularization coincide (and when not)

For ℓ linear in $\boldsymbol{\xi}$ (so $\nabla\ell = \mathbf{c}$ constant, $\mathrm{Var}_{\hat{P}_N}(\ell) = \mathbf{c}^\top\hat{\Sigma}\mathbf{c}$ for empirical covariance $\hat{\Sigma}$), the χ^2-divergence ($\phi(t) = (t-1)^2$, $\phi''(1) = 2$) penalty is $\epsilon\sqrt{2 \times 2 \times \mathbf{c}^\top\hat{\Sigma}\mathbf{c}} = 2\epsilon\sqrt{\mathbf{c}^\top\hat{\Sigma}\mathbf{c}}$, while the W_2-with-Euclidean-cost penalty is $\epsilon\,\|\mathbf{c}\|_2$. These coincide only if $\hat{\Sigma} \propto \mathbf{I}$ (isotropic data) and a specific scalar relationship between the two ϵ's holds – generically, for anisotropic $\hat{\Sigma}$, the two penalties induce *different* effective regularization directions (χ^2 penalizes $\mathbf{c}^\top\hat{\Sigma}\mathbf{c}$, a $\hat{\Sigma}$-weighted norm, while W_2/Euclidean penalizes the unweighted $\|\mathbf{c}\|_2$) – a concrete illustration that "robustness" is not a single notion: *which* ambiguity-set family one chooses materially changes *which directions* in parameter space are penalized, with real consequences for the resulting model's behaviour.

Remarks and extensions. The regularization-equivalence results of this section have an important converse direction: *any* regularizer of the form $\lambda R(\mathbf{x})$ for a norm-like R can potentially be *re-derived* as the dual of *some* ambiguity set, by "running the duality backward." This gives a way to assign a *robustness interpretation* to existing, empirically-validated regularizers, even ones originally motivated on purely statistical (e.g. Bayesian-prior) grounds – e.g. a Gaussian prior on weights (giving an ℓ_2^2 penalty in MAP estimation) corresponds, via this backward duality, to a particular (Gaussian-shaped) ambiguity set on the *labels* or *features*, providing a robustness-based complement to the standard Bayesian interpretation of ridge regression.

A further remark on adversarial training in deep learning specifically: the first-order (FGSM-type) and iterative (PGD-type) adversarial training algorithms widely used in practice correspond, respectively, to a *one-step* and *multi-step* approximation of the inner $\max_{\|\delta\|\le\epsilon} \ell(\mathbf{x}, \boldsymbol{\xi} + \delta)$ in the pointwise-WDRO formulation of this section. PGD's iterative refinement gives a tighter (less approximate) inner maximization, and correspondingly PGD-trained models tend to exhibit the gradient-norm-regularization effect of Theorem 10.5 more strongly (smaller $\|\nabla_{\boldsymbol{\xi}}\ell\|_{q^*}$) than FGSM-trained models – an empirically well-documented phenomenon that the DRO-duality lens of this Part explains as "a better approximation of the true ambiguity-set worst case yields a better-calibrated implicit regularization strength."

Theorem: training radius vs. certified radius

A model trained with WDRO radius ϵ_{train} achieves a *certified* robust radius ϵ_{cert} (largest ϵ keeping $\sup_{W_1(P,\hat{P}_N)\le\epsilon} \mathbb{E}_P[\ell(\mathbf{x}^*, \boldsymbol{\xi})]$ below a chosen threshold) equal to ϵ_{train} *only* if the threshold equals the training-time worst-case objective V^*; for any *stricter* threshold $V' < V^*$, $\epsilon_{\text{cert}} < \epsilon_{\text{train}}$.

Proof sketch. $\mathbf{x}^*$ achieves $V^* = \sup_{W_1\le\epsilon_{\text{train}}} \mathbb{E}_P[\ell(\mathbf{x}^*, \boldsymbol{\xi})]$ by definition. Since $\epsilon \mapsto \sup_{W_1\le\epsilon} \mathbb{E}_P[\ell(\mathbf{x}^*, \cdot)]$ is non-decreasing, the certified radius for threshold V^* is exactly ϵ_{train}; for $V' < V^*$, the sup reaches V' at some smaller $\epsilon < \epsilon_{\text{train}}$. ■

Example — Practical reporting implications

A model card stating "trained with $\epsilon = 0.1$ WDRO robustness" does *not* imply "certified robust to perturbations up to 0.1 at clean-accuracy level" – the certified radius at a clean-accuracy threshold could be substantially smaller than 0.1 if that threshold is stricter than the training-time worst-case loss V^*. Rigorous robustness reporting should state *both* ϵ_{train} *and* ϵ_{cert} at an explicit threshold (e.g. "90% accuracy maintained"), since the two answer different questions and, per this theorem, are not interchangeable – a distinction increasingly emphasised in ML-robustness regulatory guidance.

Historical notes. The connection between regularization and robustness has been observed repeatedly and independently across statistics and machine learning, predating its DRO formalization: Tikhonov regularization (1940s-60s, ill-posed inverse problems) was understood from early on as a form of "stabilization against perturbation," a robustness interpretation *avant la lettre*. The specific WDRO-regularization equivalence of Theorem 10.5 (Xu, Caramanis, & Mannor, 2009, for SVMs; later generalised) showed this informal connection could be made *exact* for specific (ball-uncertainty, specific loss) cases. Adversarial examples in deep learning were discovered empirically by Szegedy et al. (2013) and Goodfellow et al. (2014, introducing FGSM) well before their DRO-theoretic interpretation (Sinha et al., 2018) – another instance of empirical/applied discovery preceding theoretical unification, a pattern visible across several sections of this Part.

Theorem: equivalence class of regularization-inducing ambiguity sets for linear models

For a linear predictor $f_{\mathbf{w}}(\square) = \mathbf{w}^\top \square$ and squared loss $\ell(\mathbf{w}, \xi, y) = (y - \mathbf{w}^\top \square)^2$, *any* ambiguity set $\mathcal{P}$ whose worst-case expected loss takes the form $\sup_{Q \in \mathcal{P}} \mathbb{E}_Q[\ell] = \mathbb{E}_{\hat{P}_N}[\ell] + \kappa \|\mathbf{w}\|_q$ for some constant $\kappa > 0$ and dual exponent q induces, after minimization over $\mathbf{w}$, the *same* regularized-regression problem as ordinary ℓ_q-norm-penalized least squares with penalty weight κ – i.e. the specific *geometry* of $\mathcal{P}$ (Wasserstein ball, ϕ-divergence ball, or otherwise) is statistically irrelevant to the final estimator whenever it reduces to this penalty form, only (κ, q) matter.

Proof. If two different ambiguity sets $\mathcal{P}_1, \mathcal{P}_2$ both yield worst-case objectives of the stated form with the *same* (κ, q), the resulting optimization problems $\min_{\mathbf{w}}[\mathbb{E}_{\hat{P}_N}[\ell] + \kappa \|\mathbf{w}\|_q]$ are literally identical, hence have identical minimizers regardless of how $\mathcal{P}_1, \mathcal{P}_2$ were originally defined. ■

Example — Two different DRO derivations yielding the same ridge regression

Wasserstein-DRO with ℓ_2 ground cost on $(\square, y)$ jointly, applied to linear regression, yields (Mohajerin Esfahani–Kuhn 2018-type result) a penalty $\epsilon \|(\mathbf{w}, -1)\|_2 \approx \epsilon\sqrt{\|\mathbf{w}\|_2^2 + 1}$, which for $\|\mathbf{w}\|_2 \ll 1$ is approximately $\epsilon(1 + \frac{1}{2}\|\mathbf{w}\|_2^2)$ – an *approximate* ridge-type (ℓ_2^2) penalty. Separately, a moment-DRO formulation with an appropriately chosen covariance perturbation budget can be shown to yield an *exact* ridge penalty $\kappa \|\mathbf{w}\|_2^2$ for a specific κ determined by the perturbation budget. By the theorem (informally extended to the approximate/asymptotic regime), both routes converge to the same family of regularized estimators in the small-radius limit, confirming numerically what the general equivalence-class argument predicts: the *practical* estimator delivered by DRO, in this canonical example, depends on the ambiguity set's geometry

mainly through the resulting penalty's norm-family (ℓ_1 vs. ℓ_2 vs. other), not through finer details of which specific ambiguity-set construction produced that penalty.

Implication for adversarial training. Madry et al.'s (2018) adversarial training objective $\min_{\mathbf{w}} \mathbb{E}_{\hat{P}_N}[\max_{\|\delta\| \le \epsilon} \ell(\mathbf{w}, \xi + \delta, y)]$ is, for linear models, exactly an instance of the box/ball-uncertainty RC of Part 9 applied pointwise to each training example's feature vector – combining this section's regularization-equivalence theorem with Part 9's support-function machinery shows that, for linear models and convex losses, adversarial training's inner maximization collapses (via Eq. 9.1.ell-type identities) to an explicit norm-penalty term added to the empirical loss, recovering the gradient-penalty connection noted in this section's main text as a special case of the general equivalence-class result above – adversarial training, DRO, and norm-regularized empirical risk minimization are not three separate techniques but three derivational routes to the same (or closely related) family of estimators whenever the underlying model is linear and the loss convex.

Historical notes (addendum). The recognition that diverse robustness/regularization techniques converge to overlapping estimator families for linear models has been independently rediscovered across several literatures – robust statistics (Huber, 1960s), ridge/LASSO regression (Tikhonov, 1943; Tibshirani, 1996), and modern adversarial-ML DRO (Sinha, Namkoong, Duchi, 2018) – a recurring pattern of convergent discovery that this section's equivalence-class framing makes explicit rather than leaving as a series of separately-noted coincidences.

11.1 Knightian Uncertainty and Ambiguity Aversion

Overview

Knight (1921) distinguished *risk* (known probabilities) from *uncertainty* (unknown or partially known probabilities). *Ambiguity aversion* – the empirically robust preference for known-probability gambles over equally-likely-on-average unknown-probability gambles – motivates decision-theoretic frameworks that go beyond subjective expected utility (SEU), the formal subject of Part 11.

The Ellsberg paradox. An urn contains 90 balls: 30 red, and 60 that are black or yellow in unknown proportion. Compare gambles: A = win \$100 if red; B = win \$100 if black. Most people prefer A over B (red has known probability $30/90$; black has ambiguous probability in $[0, 60]/90$). Now compare C = win \$100 if red-or-yellow; D = win \$100 if black-or-yellow. Most people *also* prefer D over C.

Theorem: Ellsberg preferences violate SEU

No probability measure P over $\{\text{red}, \text{black}, \text{yellow}\}$ with $P(\text{red}) = 1/3$ can simultaneously satisfy $P(\text{red}) > P(\text{black})$ (from $A \succ B$) and $P(\text{black}) + P(\text{yellow}) > P(\text{red}) + P(\text{yellow})$ (from $D \succ C$, i.e. $P(\text{black}) > P(\text{red})$) – these two inequalities directly contradict each other.

Proof. $A \succ B \iff \mathbb{E}[u \cdot \mathbf{1}_{\text{red}}] > \mathbb{E}[u \cdot \mathbf{1}_{\text{black}}] \iff P(\text{red}) > P(\text{black})$ for any SEU representation with utility $u > 0$ of the prize. Similarly $D \succ C \iff P(\text{black}) + P(\text{yellow}) > P(\text{red}) + P(\text{yellow}) \iff P(\text{black}) > P(\text{red})$. The two conditions $P(\text{red}) > P(\text{black})$ and $P(\text{black}) > P(\text{red})$ cannot both hold for any single P. ■

Example — Quantifying ambiguity aversion via a capacity

Model beliefs not by a single P but by a *capacity* (non-additive measure) ν with $\nu(\text{red}) = 1/3$, $\nu(\text{black}) = \nu(\text{yellow}) = \nu(\{\text{black}, \text{yellow}\})/2$ but with $\nu(\{\text{black}, \text{yellow}\}) = 2/3$ split *pessimistically*: $\nu(\text{black}) = \nu(\text{yellow}) = \epsilon < 1/3$ each (so $\nu(\text{black}) + \nu(\text{yellow}) < 2/3 = \nu(\{\text{black}, \text{yellow}\})$, the hallmark of *superadditivity-violating*, i.e. subadditive, capacities). With $\epsilon = 0.2$: $\nu(\text{red}) = 0.333 > \nu(\text{black}) = 0.2$ gives $A \succ B$; and $\nu(\text{black}) + \nu(\text{yellow}) = 0.4 < \nu(\text{red}) + \nu(\text{yellow})$? We need $D \succ C$ i.e. a Choquet-style evaluation of $\{\text{black}, \text{yellow}\}$ exceeding that of $\{\text{red}, \text{yellow}\}$; using the capacity values directly, $\nu(\{\text{black}, \text{yellow}\}) = 2/3 > \nu(\{\text{red}, \text{yellow}\}) = \nu(\text{red}) + \epsilon = 0.333 + 0.2 = 0.533$. Indeed $2/3 \approx 0.667 > 0.533$, reproducing $D \succ C$. The formal Choquet machinery making such capacity-based evaluations rigorous is developed in Section 11.3.

Machina-Schmeidler relaxations. A weaker departure from SEU retains a single P but relaxes linearity in probabilities (rank-dependent utility), explaining some Allais-type violations. The Ellsberg paradox, however, is specifically a violation of *probabilistic sophistication itself* (Theorem 11.1), requiring the richer machinery of sets of priors (Γ-minimax, Section 11.4) or capacities (Choquet, Section 11.3).

Example — A second Ellsberg-type violation: the reflection pattern

Consider gambles on a coin of unknown bias $\theta \in [0, 1]$ vs. a known-fair coin. Gamble E pays on heads of the fair coin; F on heads of the unknown coin. Many prefer $E \succ F$. For tails, $E' \succ F'$ similarly. Under SEU with subjective θ, $E \succ F \iff 1/2 > \theta$ and $E' \succ F' \iff \theta > 1/2$ – a direct contradiction, structurally identical to Theorem 11.1's proof but in a different setting, confirming ambiguity aversion is a robust, replicated phenomenon.

Smooth ambiguity preferences. Klibanoff-Marinacci-Mukerji (2005) propose $V = \mathbb{E}_\mu[\phi(\mathbb{E}_\pi[u])]$ where μ is a second-order prior over first-order priors π and ϕ is concave. This is a special case of Γ-minimax (Section 11.4) in the limit of steep ϕ, but for smooth ϕ gives interior-weighted (not worst-case-only) ambiguity attitudes – a third formal language for the same phenomenon.

Implications. The Ellsberg paradox is the decision-theoretic analogue of the geometric uncertainty sets of Part 9: just as a single point estimate $\bar{\mathbf{a}}$ is replaced by a set $\mathcal{U}$, a single probability measure P is replaced by a *set of measures* or a non-additive capacity ν. The remainder of Part 11 develops formal axiomatizations (maximin EU, Choquet EU, Γ-minimax, info-gap) of preferences consistent with ambiguity aversion, each corresponding – as will become apparent – to a different geometric or measure-theoretic structure on the "ambiguity set of beliefs," echoing the uncertainty-set taxonomy of Section 9.1 and the ambiguity-set taxonomy of Section 10.1.

Theorem: no finite set of pairwise comparisons can fully reveal a non-additive capacity

For $|\Theta| = n$, an additive probability P has $n - 1$ free parameters, fully determined by $n - 1$ independent comparisons (e.g. via standard probability-elicitation methods). A general capacity ν has $2^n - 2$ free parameters. For $n \geq 3$, $2^n - 2 > n - 1$ (e.g. $n = 3$: $6 > 2$), so *strictly more* pairwise/few-outcome comparisons are needed to elicit ν than P – and the *number of comparisons needed grows exponentially in n for ν vs. linearly for P*.

Proof. Immediate from counting free parameters: ν assigns a value in $[0, 1]$ to each of the 2^n

subsets of Θ, minus 2 fixed values ($\nu(\emptyset) = 0, \nu(\Theta) = 1$), giving $2^n - 2$ free parameters; P assigns n values summing to 1, giving $n - 1$ free parameters. For $n \geq 3$: $2^n - 2 - (n-1) = 2^n - n - 1 > 0$ for $n \geq 3$ (check: $n = 3$, $8 - 4 = 4 > 0$). ■

Example — The elicitation burden in practice

For $|\Theta| = 10$ (e.g. 10 macroeconomic scenarios), a full capacity has $2^{10} - 2 = 1022$ free parameters – entirely infeasible to elicit directly from a decision-maker (compare $n - 1 = 9$ for an additive P). This is *why*, in practice, capacity-based (Choquet) and prior-set-based (Γ-minimax) methods almost always use *parametric* families – ϵ-contamination (1 extra parameter beyond π_0), Hurwicz-type capacities (1 parameter α), or belief functions built from a small number of "focal elements" (e.g. ≤ 10 sets with non-zero Möbius mass, rather than all 2^{10}) – as already flagged in Section 11.3's remarks. This theorem quantifies *why* that practical restriction is essentially unavoidable for any $|\Theta|$ beyond the smallest toy examples, and explains why real-world applications of CEU/Γ-minimax invariably involve such parametric restrictions rather than "fully general" capacities or prior sets.

Remarks and extensions. Beyond the Ellsberg urn and the reflection-pattern examples, ambiguity aversion has been documented in a wide range of real-world contexts: insurance markets (people pay a premium to avoid ambiguous, rather than merely risky, perils – "ambiguity premia" distinct from standard risk premia), financial markets (the "home bias" puzzle, where investors overweight domestic assets whose distributions feel less ambiguous than foreign assets', even when historical risk-return profiles are similar), and medical decision-making (patients and physicians often prefer treatments with well-characterised (even if higher) risk profiles over novel treatments with ambiguous risk profiles, a pattern directly relevant to the kind of robust decision-making frameworks developed in this Part).

A historical remark: although Ellsberg's 1961 paper crystallised the formal paradox, the underlying intuition – that "unknown unknowns" warrant different treatment than "known unknowns" – traces back at least to Keynes (1921) and, as noted at the start of this section, Knight (1921), both writing in the aftermath of WWI in a period of acute economic uncertainty. The formal decision-theoretic frameworks of Sections 11.2–11.5 can be read as successive attempts, across nearly a century, to give Knight's informal distinction precise mathematical content – a long arc that mirrors, on a smaller scale, this volume's own progression from informal robustness intuitions (Part 9) to fully axiomatized frameworks (Part 11).

Theorem: ambiguity aversion under repeated stationary sampling

For an ambiguous urn drawn from N times i.i.d. (fixed but unknown composition across all draws), the per-draw certainty-equivalent value under any reasonable Γ-minimax rule converges, as $N \to \infty$, to the value under the *true* composition – ambiguity aversion's effect vanishes in the large-N limit for a fixed, learnable parameter, despite having full force at $N = 1$.

Proof sketch. As $N \to \infty$, the empirical composition converges a.s. (SLLN) to the true composition, shrinking any reasonable Γ (e.g. a Bayesian credible set updated on observed draws) to a point mass. Γ-minimax over a shrinking-to-a-point Γ converges to ordinary expected utility under that point. ■

Example — When ambiguity aversion matters most

The pandemic-policy case study involves a genuinely *one-shot* decision (a novel pathogen, not repeatable). The theorem suggests ambiguity aversion's practical impact *should* diminish for *repeated, stationary* decisions as evidence accumulates – remaining maximally ambiguity-averse after thousands of informative draws is, technically, "failing to learn." This gives a normative criterion for when ambiguity-averse frameworks (Sections 11.2-11.5) are most appropriate: genuinely novel, non-repeatable, or non-stationary decisions, rather than routine repeated ones where Bayesian updating is the appropriate – and ultimately ambiguity-*neutral* – long-run framework.

Historical notes. Daniel Ellsberg's 1961 paper ("Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*') introduced the urn experiments of this section as a direct challenge to Leonard Savage's then-recent (1954) axiomatization of subjective expected utility – itself the culmination of a research program tracing through von Neumann & Morgenstern (1944, expected utility under *known* probabilities) and Ramsey/de Finetti (1920s-30s, subjective probability). Ellsberg's own later career (as a Pentagon analyst and, famously, the source of the 1971 Pentagon Papers leak) is sometimes noted as a curious historical footnote, though it bears no direct mathematical relation to his decision-theory work. The Knight (1921) risk/uncertainty distinction predates Ellsberg by four decades but lacked, until Ellsberg's experiments, a sharp empirical demonstration that the distinction had *behavioural* (not merely conceptual) force – the gap between conceptual distinction and formal axiomatization that this entire Part traces.

Theorem: the Ellsberg paradox as a violation of probabilistic sophistication

No single subjective probability measure P over the "unknown urn" composition in the classic Ellsberg (1961) two-urn experiment can simultaneously rationalize the modally observed preference pattern (preferring the known-risk bet over the ambiguous bet in *both* the "bet on red" and "bet on black" framings) under expected-utility maximization with any non-degenerate utility function – the observed choices are inconsistent with *any* probabilistic-sophistication-respecting model, not merely with a specific (e.g. risk-neutral) one.

Proof sketch. Suppose a subjective probability $p = \Pr[\text{red in ambiguous urn}]$ rationalized both observed preferences. Preferring the known-risk "bet on red" (urn with exactly 50% red) over the ambiguous urn's "bet on red" requires (under expected utility) $p < 0.5$; but by an entirely symmetric argument, preferring the known-risk "bet on black" over the ambiguous urn's "bet on black" requires $\Pr[\text{black}] = 1 - p < 0.5$, i.e. $p > 0.5$ – a direct contradiction, so no single p can rationalize both observed preferences. ■

Example — Quantifying ambiguity aversion via the Ellsberg gap

A common experimental quantification defines the *Ellsberg gap* as the difference between the certainty-equivalent of the known-risk bet and the certainty-equivalent of the ambiguous bet, both nominally offering the same 50%-type payout structure. Typical experimental gaps (across replications of Ellsberg-type designs) range from 5-15% of the bet's face value – i.e. subjects typically demand a 5-15% discount to accept ambiguity over equivalent quantified risk, a magnitude commensurate with, though generally somewhat smaller than, typical risk premia for genuine (quantified) risk in similar stakes ranges, underscoring that ambiguity aversion and risk aversion are empirically

distinct phenomena of comparable (if not always equal) behavioural importance.

Connection to the multiple-priors framework. The Ellsberg paradox's resolution via Gilboa–Schmeidler's (1989) maximin expected utility (Section 11.4's Γ-minimax) is direct: representing the ambiguous urn's uncertainty by the full set $\Gamma = \{P : \Pr[\text{red}] \in [0,1]\}$ (complete ignorance) rather than a single P, and evaluating bets via $\inf_{P\in\Gamma} \mathbb{E}_P[\text{payoff}]$, immediately rationalizes preferring the known 50% urn over the ambiguous urn for *both* colour framings simultaneously (since $\inf_{P\in\Gamma} \Pr[\text{red}] = 0$ and $\inf_{P\in\Gamma} \Pr[\text{black}] = 0$, both strictly less than the known urn's 0.5) – the very feature (asymmetric treatment of both framings) that Theorem 11.1 shows is impossible for any single-prior expected-utility model becomes immediate once the decision criterion itself is built on a *set* of priors rather than one.

Historical notes (addendum). While Ellsberg's (1961) original paper presented the paradox primarily as an empirical challenge to Savage's (1954) subjective expected utility axioms, its formal resolution via an explicit multiple-priors decision rule had to await Gilboa and Schmeidler's (1989) axiomatization, nearly three decades later – a notable example, recurring throughout this Part's historical notes, of empirical anomalies preceding their formal axiomatic resolution by a substantial gap, mirroring the analogous multi-decade gap between Knight's (1921) informal uncertainty/risk distinction and its formal multiple-priors mathematization.

11.2 Maximin, Minimax Regret, and Hurwicz Criteria

Overview

The simplest formal responses to Knightian uncertainty replace expected utility $\mathbb{E}_P[u]$ by a worst-case or regret-based criterion. This section formalises three classical decision criteria – maximin (Wald), minimax regret (Savage), and the Hurwicz α-criterion – as functionals on the set of possible states, and characterises their axiomatic relationships.

Definition: The three criteria

Let $\mathcal{A}$ be a set of actions, Θ a set of states, and $u(a,\theta)$ the payoff of action a in state θ. Define:

$$\text{Maximin (Wald):} \quad a^* = \arg\max_{a\in\mathcal{A}} \min_{\theta\in\Theta} u(a,\theta)$$

$$\text{Minimax regret (Savage):} \quad a^* = \arg\min_{a\in\mathcal{A}} \max_{\theta\in\Theta} \left[\max_{a'\in\mathcal{A}} u(a',\theta) - u(a,\theta)\right]$$

$$\text{Hurwicz}(\alpha): \quad a^* = \arg\max_{a\in\mathcal{A}} \Big[\alpha \min_{\theta} u(a,\theta) + (1-\alpha)\max_{\theta} u(a,\theta)\Big], \ \alpha\in[0,1]$$

Relationships. Maximin is the special case $\alpha = 1$ of Hurwicz; pure optimism ($\alpha = 0$) recovers $\arg\max_a \max_\theta u(a,\theta)$. Minimax regret is generally *not* a special case of Hurwicz, since regret is defined relative to the column-maximum $\max_{a'} u(a',\theta)$, a quantity that depends on the whole action set $\mathcal{A}$ – minimax regret can therefore violate the *independence of irrelevant alternatives*, while Hurwicz criteria (for fixed α) do not.

Example — A 3×3 decision table

Payoff matrix $u(a, \theta)$, rows = actions, columns = states:

	θ_1	θ_2	θ_3
a_1	10	2	4
a_2	5	6	5
a_3	0	1	12

Maximin: row minima are $2, 5, 1$ for a_1, a_2, a_3; $\max(2, 5, 1) = 5$, so $a^* = a_2$.
Minimax regret: column maxima are $10, 6, 12$. Regret table: a_1: $(0, 4, 8)$; a_2: $(5, 0, 7)$; a_3: $(10, 5, 0)$. Max regrets: $8, 7, 10$. $\min(8, 7, 10) = 7$, so $a^* = a_2$ (here both criteria agree, but this is not generic).
Hurwicz with $\alpha = 0.5$: row scores $0.5 \cdot \min + 0.5 \cdot \max$: a_1: $0.5(2) + 0.5(10) = 6$; a_2: $0.5(5) + 0.5(6) = 5.5$; a_3: $0.5(1) + 0.5(12) = 6.5$. $\max = 6.5$, so $a^* = a_3$ – the optimistic weighting flips the choice to the action with the highest *best-case* payoff, illustrating sensitivity to α.

Axiomatic characterization of maximin. Maximin preferences satisfy Weak Order, Monotonicity, and Certainty-Independence, but violate Savage's Sure-Thing Principle under ambiguity. Gilboa-Schmeidler (1989) show maximin over a set of priors Γ (Γ-minimax, Section 11.4) is the *unique* representation satisfying these axioms plus continuity – so "maximin over Γ" is axiomatically forced, not ad-hoc.

Example — A 4×2 table illustrating regret's action-set dependence

Payoffs: $a_1 : (8, 2)$, $a_2 : (5, 5)$, $a_3 : (2, 8)$. Column maxima $(8, 8)$. Regrets: $a_1 : (0, 6)$, $a_2 : (3, 3)$, $a_3 : (6, 0)$. Max regrets $6, 3, 6$ – minimax regret picks a_2. Now add $a_4 : (9, 0)$. New column maxima $(9, 8)$. a_2's regret becomes $(4, 3)$, max regret $4 > 3$ – purely due to a_4's addition, an instance of regret's dependence on the full action set.

Hurwicz and optimism calibration. $\alpha \in [0, 1]$ can be interpreted as a measure of optimism/pessimism, elicitable from choices over two-outcome lotteries, analogous to risk-aversion elicitation. However, because Hurwicz collapses all intermediate states to irrelevance, it is informationally coarser than Choquet (Section 11.3) or Γ-minimax (Section 11.4), which use the entire payoff vector.

Discussion. Maximin and minimax-regret are both "extreme pessimism" criteria in the sense of placing all decision weight on the worst (or most-regretted) state – formally, they are the Γ-minimax criterion of Section 11.4 with Γ equal to the set of *all* probability distributions over Θ (point masses included), the degenerate extreme of the ambiguity-set hierarchy. The Hurwicz criterion interpolates toward standard expected-utility-like averaging but with only two "effective states" (best and worst), losing the rich distributional information used by the Choquet and robust-Bayesian criteria of Sections 11.3–11.4. All three criteria, however, share the key formal property of being *not* expectations with respect to any single additive probability measure – the technical content of the impossibility result of Section 11.1, generalised axiomatically in Section 11.3.

Theorem: Hurwicz criterion is the unique criterion depending only on $(\min, \max)$ satisfying monotonicity and continuity

Any decision criterion $V(a) = F(\min_\theta u(a,\theta), \max_\theta u(a,\theta))$ for a continuous, monotone-increasing-in-both-arguments $F : \mathbb{R}^2 \to \mathbb{R}$ with $F(t,t) = t$ for all t (normalization: constant payoffs are valued at their constant value) can be written as $V(a) = \alpha \min_\theta u(a,\theta) + (1-\alpha)\max_\theta u(a,\theta)$ for some $\alpha \in [0,1]$ depending only on F (not on a or Θ), *provided* F is additionally required to be *linear* along the diagonal-normalized directions – i.e. the Hurwicz form is the unique *affine* such F; without linearity, other monotone $(\min, \max)$-based criteria (e.g. $F(t_1,t_2) = \sqrt{t_1 t_2}$ for $t_1, t_2 \geq 0$, a "geometric Hurwicz") also satisfy monotonicity, continuity, and the normalization.

Proof sketch. Linearity plus the normalization $F(t,t) = t$ forces $F(t_1,t_2) = \alpha t_1 + (1-\alpha)t_2$ for $\alpha = 1 - F_2(t,t)$ (partial derivative, constant by linearity), with $\alpha \in [0,1]$ by monotonicity (F increasing in both arguments requires $\alpha \geq 0$ and $1-\alpha \geq 0$). Without linearity, e.g. $F(t_1,t_2) = \sqrt{t_1 t_2}$ satisfies $F(t,t) = t$, is monotone (for $t_1, t_2 > 0$), and continuous, but is not of the Hurwicz form. ■

Example — The geometric-Hurwicz alternative, numerically

For the Exercise-11.2.1-style matrix with action a_1 having $(\min, \max) = (2, 10)$: standard Hurwicz($\alpha = 0.5$) gives $0.5(2) + 0.5(10) = 6$. Geometric-Hurwicz gives $\sqrt{2 \times 10} = \sqrt{20} \approx 4.47$ – a *lower* (more pessimistic) value than the arithmetic version for the same "$\alpha = 0.5$-like" symmetric treatment of min and max, because the geometric mean is always $\leq$ the arithmetic mean (AM-GM inequality), with equality only when min $=$ max (no uncertainty). This shows that "which F" is itself a substantive modelling choice with real consequences – the geometric form penalizes *spread* between min and max more aggressively than the arithmetic (Hurwicz) form, an additional, rarely-discussed degree of freedom in classical decision criteria beyond the single parameter α.

Remarks and extensions. The three criteria of this section (maximin, minimax regret, Hurwicz) are sometimes collectively called "classical" decision criteria to distinguish them from the more recent (1980s onward) axiomatically-grounded frameworks of Sections 11.3–11.4. It is worth emphasising that "classical" here does not mean "outdated" – maximin in particular remains the dominant practical criterion in many engineering safety-critical contexts (e.g. structural engineering safety factors are, in effect, maximin criteria over a box-type uncertainty set for material properties, directly connecting back to Section 9.1's box uncertainty), precisely *because* of its conceptual simplicity and the ease with which $\min_\theta u(a,\theta)$ can be communicated to non-specialist stakeholders (regulators, the public) compared to, say, a Γ-minimax Bayes risk.

A computational remark: for large action sets $\mathcal{A}$ and state sets Θ, even maximin's $\max_a \min_\theta$ can be expensive if $u(a,\theta)$ must be evaluated by simulation for each (a,θ) pair. *Bayesian optimization over the worst case* – using a surrogate model (e.g. a Gaussian process) for $u(a,\theta)$ and actively selecting (a,θ) pairs to evaluate, prioritising regions likely to contain either the maximin optimum or a currently-underestimated worst case – is an active area combining the decision-theoretic content of this section with the surrogate-modelling techniques covered in the companion series' chapters on Bayesian optimization.

Theorem: relative regret mitigates but does not eliminate IIA violation

Define relative regret $R(a,\theta) = \frac{\max_{a'} u(a',\theta) - u(a,\theta)}{\max_{a'} u(a',\theta) - \min_{a'} u(a',\theta)}$. Minimizing $\max_\theta R(a,\theta)$ is *less* sensitive to adding new alternatives than absolute minimax regret when the new alternative scales the per-state range proportionally, but can still violate strict independence-of-irrelevant-alternatives when a new alternative shifts only the max (not the min) for some state.

Proof sketch. If a new action a_4 shifts both $\max_{a'} u(a',\theta)$ and $\min_{a'} u(a',\theta)$ proportionally, the range max $-$ min scales correspondingly and other actions' relative regrets are unchanged. If a_4 shifts only the max (e.g. $a_4 = (9,0)$ raising column-1's max from 8 to 9 without changing its min), the range for column 1 changes from 6 to 7 (using min$= 2$) or similar, so other actions' relative regrets still shift, just by a smaller proportional amount than absolute regret. ■

Example — Relative regret for the running 3×2 example

Matrix $a_1 = (8,2), a_2 = (5,5), a_3 = (2,8)$; column maxima $(8,8)$, minima $(2,2)$, ranges $(6,6)$. Relative regrets: $a_1 : (0,1)$, $a_2 : (0.5, 0.5)$, $a_3 : (1,0)$; minimax relative regret picks a_2 (0.5), matching absolute regret here. Adding $a_4 = (9,0)$: column maxima $(9,8)$, minima $(2,0)$, ranges $(7,8)$. a_2's relative regrets become $((9-5)/7, (8-5)/8) = (0.571, 0.375)$, max 0.571 (up from 0.5) – a smaller proportional shift than absolute regret's jump from 3 to 4 for less symmetric examples, though here the relative changes happen to be comparable.

Historical notes. Abraham Wald's maximin criterion (1945, "Statistical Decision Functions Which Minimize the Maximum Risk") predates Ellsberg by 16 years and was developed within *statistical decision theory* (choosing estimators/tests to minimize worst-case risk over a parameter space) rather than in response to any ambiguity-aversion puzzle – Wald's motivation was purely about statistical estimation under an unknown true parameter, with no behavioural-economics content. Savage's own 1951 minimax-regret criterion, and Hurwicz's 1951 α-criterion, were part of the same early-1950s statistical-decision-theory ferment (largely centred at the Cowles Commission and RAND Corporation, reflecting the Cold War's influence on decision theory's early development – worst-case thinking had obvious appeal for nuclear-strategy-adjacent research). The retrospective recognition that these 1940s-50s statistical-decision criteria *also* resolve (or at least formalize) the 1961 Ellsberg paradox is a later (1980s onward) synthesis, again illustrating cross-temporal convergence of independently-motivated research programs.

A puzzle to motivate the result. Two actions have payoffs $\mathbf{x}_1 = (10,10,10)$ and $\mathbf{x}_2 = (5,15,8)$ across three equally plausible, unknown states. Since $\mathbf{x}_2$ beats $\mathbf{x}_1$ in one state and loses in two, the Hurwicz ranking between them genuinely depends on the optimism parameter α: $H_0(\mathbf{x}_1) = \min = 10 > H_0(\mathbf{x}_2) = \min = 5$ (the fully pessimistic criterion, $\alpha = 0$, favours $\mathbf{x}_1$), while $H_1(\mathbf{x}_1) = \max = 10 < H_1(\mathbf{x}_2) = \max = 15$ (the fully optimistic criterion, $\alpha = 1$, favours $\mathbf{x}_2$) – the ranking flips somewhere in between. Setting $H_\alpha(\mathbf{x}_1) = H_\alpha(\mathbf{x}_2)$: $10 = \alpha(15) + (1-\alpha)(5) = 5 + 10\alpha \implies \alpha = 0.5$ is the crossover point; for $\alpha < 0.5$, $\mathbf{x}_1$ wins, for $\alpha > 0.5$, $\mathbf{x}_2$ does. This α-sensitivity is exactly the criticism most often levelled at Hurwicz's criterion – but is it *always* this unstable, or only when the two actions are genuinely incomparable in some sense?

Theorem: the Hurwicz criterion's optimism parameter and first-order stochastic dominance

For the Hurwicz criterion $H_\alpha(\mathbf{x}) = \alpha \max_\theta u(\mathbf{x}, \theta) + (1-\alpha) \min_\theta u(\mathbf{x}, \theta)$, if action $\mathbf{x}_1$'s payoff distribution (over the unknown state θ, treated as if uniformly weighted for this comparison) first-order stochastically dominates $\mathbf{x}_2$'s, then $H_\alpha(\mathbf{x}_1) \geq H_\alpha(\mathbf{x}_2)$ for *every* $\alpha \in [0, 1]$ – no crossover of the kind seen above is possible once one action dominates the other, regardless of which α is chosen.

Proof. First-order stochastic dominance of $\mathbf{x}_1$ over $\mathbf{x}_2$ implies $u(\mathbf{x}_1, \theta) \geq u(\mathbf{x}_2, \theta)$ for every θ after an appropriate state-relabelling consistent with the dominance ordering (a standard equivalent characterization of FOSD for finite state spaces). Pointwise dominance immediately implies dominance of both the max and the min, hence of any convex combination H_α. ■

So the puzzle's instability is, in retrospect, exactly diagnostic: the two payoff vectors above admit no FOSD relationship (neither dominates pointwise after any relabelling, since $\mathbf{x}_2$ exceeds $\mathbf{x}_1$ in one state and falls short in the other two), which is precisely why the theorem's conclusion does not apply and a genuine α-dependence emerges. Whenever a clean dominance relationship *does* hold between two candidate actions, Hurwicz's notorious parameter-sensitivity simply does not arise – a useful first check before worrying about which α to use.

Practical use of the crossover computation. More generally, reporting not just a single Hurwicz ranking (for one arbitrarily chosen α) but the full crossover structure across $\alpha \in [0, 1]$ – as computed above for the motivating example – gives a decision-maker a richer, assumption-transparent picture than a single point estimate, and is the Hurwicz-criterion analogue of the sensitivity-to-Γ analysis recommended for Bertsimas–Sim budgets in Section 9.4.

An open question. The theorem gives a sufficient condition (FOSD) for α-independence of the ranking, but it is not necessary: two non-dominating payoff vectors can still happen to rank consistently across all $\alpha \in [0, 1]$ if their crossover point (computed as above) falls outside $[0, 1]$ entirely. Characterizing exactly which pairs of non-dominating actions nonetheless yield an α-stable ranking – beyond simply computing the crossover point case by case – does not appear to have a clean closed-form answer in the existing decision-theory literature, and is left here as a natural question for the interested reader.

11.3 Choquet Expected Utility and Non-Additive Measures

Overview

Choquet Expected Utility (CEU; Schmeidler, 1989) generalises SEU by replacing the additive probability measure P with a non-additive *capacity* ν and replacing the ordinary (Lebesgue/Riemann) integral with the *Choquet integral*, recovering exactly the kind of pessimistic capacities sketched informally in Section 11.1's Ellsberg example.

Definition: Capacity and Choquet integral

A *capacity* on $(\Theta, \mathcal{F})$ is a set function $\nu : \mathcal{F} \to [0, 1]$ with $\nu(\emptyset) = 0$, $\nu(\Theta) = 1$, and *monotonicity* $A \subseteq B \Rightarrow \nu(A) \leq \nu(B)$ (note: *not* required to be additive). For a bounded

measurable $f : \Theta \to \mathbb{R}_{\geq 0}$, the Choquet integral is

$$\int f \, d\nu \triangleq \int_0^\infty \nu(\{\theta : f(\theta) \geq t\}) \, dt.$$

If ν is additive (a true probability measure), this coincides with the ordinary expectation $\mathbb{E}_\nu[f]$.

Theorem (CEU representation of ambiguity-averse preferences). If ν is *subadditive* – $\nu(A \cup B) \leq \nu(A) + \nu(B)$ for disjoint A, B – then the Choquet integral $\int u(a, \theta) \, d\nu(\theta)$ assigns systematically lower value to "ambiguous" acts (those whose payoff varies a lot across states not individually assessed) than the average-case additive expectation would, formally capturing ambiguity aversion. Conversely, *superadditive* ν ($\nu(A \cup B) \geq \nu(A) + \nu(B)$) represents ambiguity *seeking*.

Proof sketch (*subadditive case, finite* $\Theta = \{\theta_1, \dots, \theta_n\}$, $u(a, \theta_1) \leq \cdots \leq u(a, \theta_n)$). The Choquet integral has the explicit finite-sum form

$$\int u \, d\nu = u(\theta_1) + \sum_{k=2}^{n} [u(\theta_k) - u(\theta_{k-1})] \, \nu(\{\theta_k, \dots, \theta_n\}).$$

Each increment $[u(\theta_k) - u(\theta_{k-1})] \geq 0$ is weighted by ν of the *upper tail* $\{\theta_k, \dots, \theta_n\}$. Subadditivity implies $\nu(\{\theta_k, \dots, \theta_n\}) \leq \sum_{j \geq k} \nu(\{\theta_j\})$ can be *less* than the additive-measure weight on the same tail under a "central" P with the same singleton values, so high payoffs (top of the sorted list) get systematically down-weighted relative to SEU – a pessimistic bias. ■

Example — Computing a Choquet integral

$\Theta = \{\theta_1, \theta_2, \theta_3\}$, $u = (2, 5, 9)$ (already sorted). Take the subadditive capacity $\nu(\{\theta_1\}) = 0.2, \nu(\{\theta_2\}) = 0.2, \nu(\{\theta_3\}) = 0.2, \nu(\{\theta_2, \theta_3\}) = 0.5, \nu(\{\theta_1, \theta_2, \theta_3\}) = 1$ (check subadditivity: $\nu(\{\theta_2, \theta_3\}) = 0.5 < \nu(\{\theta_2\}) + \nu(\{\theta_3\}) = 0.4$? No, $0.5 > 0.4$ – so this particular ν is actually *super*additive on this pair; for a genuinely subadditive example take $\nu(\{\theta_2, \theta_3\}) = 0.3$). With $\nu(\{\theta_2, \theta_3\}) = 0.3$: $\int u \, d\nu = u(\theta_1) + [u(\theta_2) - u(\theta_1)]\nu(\{\theta_2, \theta_3\}) + [u(\theta_3) - u(\theta_2)]\nu(\{\theta_3\}) = 2 + (5-2)(0.3) + (9-5)(0.2) = 2 + 0.9 + 0.8 = 3.7$. Compare the additive expectation under $P(\theta_i) = 1/3$: $\mathbb{E}_P[u] = (2+5+9)/3 = 5.33$. The Choquet value 3.7 is substantially below 5.33, quantifying the pessimistic adjustment.

Möbius transform and the core of a capacity. Every capacity ν admits a Möbius transform $m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} \nu(B)$. The *core* $\text{core}(\nu) = \{P : P(A) \geq \nu(A) \ \forall A\}$ satisfies $\int f \, d\nu = \min_{P \in \text{core}(\nu)} \mathbb{E}_P[f]$: **the Choquet integral w.r.t. subadditive ν equals the Γ-minimax Bayes risk with $\Gamma = \text{core}(\nu)$** – the bridge between Section 11.3 and 11.4.

Example — The core of the Section 11.3 capacity

With $\nu(\{\theta_i\}) = 0.2$ each, $\nu(\{\theta_2, \theta_3\}) = 0.3$, $\nu(\{\theta_1, \theta_2\}) = \nu(\{\theta_1, \theta_3\}) = 0.4$, $\nu(\Theta) = 1$, the core requires $P(\theta_i) \geq 0.2$, $P(\theta_2) + P(\theta_3) \geq 0.3$, $P(\theta_1) + P(\theta_2) \geq 0.4$, $P(\theta_1) + P(\theta_3) \geq 0.4$. The vertex $P = (0.4, 0.3, 0.3)$ satisfies several with equality. Minimizing $\mathbb{E}_P[u]$ for $u = (2, 5, 9)$ over this polytope gives 3.7, matching the Choquet value – an alternative computational route.

Dempster-Shafer belief functions. Belief functions, $\nu(A) = \sum_{B \subseteq A} m(B)$ for $m \geq 0$, $\sum_B m(B) = 1$, are automatically subadditive (totally monotone), so their Choquet integrals coincide

with Γ-minimax over a non-empty core – a formalization of "evidence consistent with, but not pinning down, a single probability."

Connections. CEU subsumes maximin (the capacity $\nu(A) = 0$ for all $A \neq \Theta$, $\nu(\Theta) = 1$ gives $\int u\, d\nu = \min_\theta u(\theta)$) and ordinary SEU (additive $\nu = P$) as extreme special cases, providing the formal unification promised at the end of Section 11.2. The next section, Γ-minimax, offers an alternative – and in many applications more tractable – generalisation, replacing the single capacity ν with an explicit *set* Γ of additive priors, connecting back to the set-based (rather than capacity-based) language of Parts 9–10.

Theorem: Choquet integral of a monotone transformation

For a strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$ and capacity ν, $\int \psi(u(\theta))\, d\nu(\theta) \neq \psi\left(\int u(\theta)\, d\nu(\theta)\right)$ in general (the Choquet integral does not commute with monotone transformations of the integrand) – in sharp contrast to the additive case where, of course, neither side equals the other either, but the *ordering* of acts by $\int \psi(u)\, d\nu$ vs. $\psi(\int u\, d\nu)$ can *differ* for the Choquet integral in a way it cannot for additive $\mathbb{E}_P$ (where $\mathbb{E}_P[\psi(u)]$ and $\psi(\mathbb{E}_P[u])$ relate via Jensen's inequality with a *fixed* direction depending only on ψ's convexity, regardless of the act u).

Proof sketch (*counterexample suffices*). Take $\Theta = \{\theta_1, \theta_2\}$, $\nu(\{\theta_1\}) = \nu(\{\theta_2\}) = 0.3$ (subadditive), $u_A = (1, 4)$, $u_B = (2, 2)$ (constant). $\int u_A\, d\nu = 1 + (4 - 1)(0.3) = 1.9$ (using the finite-sum Choquet formula). $\int u_B\, d\nu = 2$ (constant act, Choquet integral of a constant equals the constant regardless of ν, since $\nu(\Theta) = 1$). So $u_B \succ u_A$ under $\int \cdot\, d\nu$ ($2 > 1.9$). Now let $\psi(t) = t^2$: $\psi(u_A) = (1, 16)$, $\int \psi(u_A)\, d\nu = 1 + (16 - 1)(0.3) = 5.5$. $\psi(u_B) = (4, 4)$, $\int \psi(u_B)\, d\nu = 4$. Now $u_A \succ u_B$ under $\int \psi(\cdot)\, d\nu$ ($5.5 > 4$) – the ranking *reversed* under the monotone transformation $\psi(t) = t^2$, something that cannot happen for additive $\mathbb{E}_P$ and any fixed u_A, u_B if ψ is monotone (monotone transformations preserve the *pointwise* ordering $u_A(\theta)$ vs. $u_B(\theta)$ at each θ, hence preserve $\mathbb{E}_P$-rankings whenever one act pointwise-dominates... but here neither dominates, so even for $\mathbb{E}_P$ a reversal is *possible* – the genuinely Choquet-specific phenomenon is that the reversal can occur even when, as here, u_B is *constant*, for which $\mathbb{E}_P[u_B] = \psi(\mathbb{E}_P[u_B])$ trivially commutes, yet the Choquet integral of $\psi(u_B)$ still differs in its relation to $\int \psi(u_A) d\nu$ in a ν-dependent, non-trivial way). ■

Example — Practical implication: utility-function choice matters more under CEU

The reversal above means that, under CEU, *which utility function* $u(\cdot)$ (*or* $\psi(\cdot)$) *is used to represent payoffs* can change *rankings* of acts, not merely *how much* one act is preferred to another (as in standard expected-utility theory, where utility-function choice affects risk-aversion *degree* but, for a fixed probability P, cannot reverse the ranking of two acts that are both non-constant unless ψ is non-monotone). Practitioners using CEU/Choquet-based robust decision criteria must therefore pay close attention to the *units/scale* in which payoffs are expressed (linear dollars? log-dollars? some other utility scale?) *before* applying the capacity ν – a methodological subtlety with no analogue in the additive-probability case, and one not always appreciated in applied CEU work.

Remarks and extensions. The Choquet integral, beyond its role in CEU, appears throughout mathematics under different names: in cooperative game theory, the Choquet integral of a capacity (there called a "characteristic function") with respect to a permutation gives the *Shapley value* contribution along that permutation, and the *average* Choquet integral over all permutations *is* the Shapley value – a deep connection between ambiguity-averse decision theory and fair-division/cost-allocation theory. Readers encountering Shapley values

in the context of feature-importance methods for machine learning (a major topic in the interpretability chapters of the companion Neural Networks series) are, in effect, encountering a disguised instance of the Choquet-integral machinery of this section.

A further remark on *elicitation*: capacities ν, having $2^{|\Theta|}-2$ free parameters (one for each non-trivial subset of Θ, modulo $\nu(\emptyset)=0, \nu(\Theta)=1$), are far harder to elicit from a decision-maker than a single additive P (with only $|\Theta|-1$ free parameters). In practice, *parametric* families of capacities – e.g. the Hurwicz-type capacities of Section 11.2 (only 1 extra parameter α), or capacities derived from a single "ambiguity aversion coefficient" applied multiplicatively to a base P_0 – are used to keep elicitation tractable, at the cost of restricting to a (small) sub-family of all possible CEU preferences.

Theorem: CEU and rank-dependent utility coincide for real-valued acts

For $\mathbb{R}$-valued acts $u:\Theta\to\mathbb{R}$, the Choquet integral $\int u\,d\nu$ coincides with the rank-dependent utility functional $\mathrm{RDU}(u)=\sum_k u_{(k)}[w(F(u_{(k)}))-w(F(u_{(k-1)}))]$ for a probability-weighting function w derived from ν via $w(F(u_{(k)}))=\nu(\{\theta:u(\theta)\geq u_{(k)}\})$ – CEU and RDU are mathematically identical for single real-valued payoffs.

Proof sketch. Both formulas are weighted sums of sorted payoffs with weights determined by a monotone transformation of cumulative upper-level-set measures; setting $w\circ F$ equal to ν on upper-level sets makes the formulas identical term-by-term. ■

Example — Prospect theory's w as an implicit capacity

Kahneman-Tversky prospect theory uses an empirically-calibrated inverse-S-shaped $w(p)$ (overweighting small probabilities, underweighting large ones). By the theorem, this corresponds to an implicit capacity $\nu(\{u\geq\text{median}\})=w(0.5)\approx 0.4<0.5$ – a *subadditive* capacity (since $\nu(A)+\nu(A^c)\approx 0.8<1$ for $A=\{u\geq\text{median}\}$), matching the pessimism pattern of Section 11.3's capacities, but here derived from *risk* (known-P) experiments rather than ambiguity experiments. This identity means decades of $w(\cdot)$-calibration studies from the risk literature could, in principle, inform capacity calibration for CEU-based ambiguity models – an underexploited cross-fertilisation.

Historical notes. The Choquet integral is named for Gustave Choquet, who introduced *capacities* in 1953 in the context of potential theory (a branch of analysis related to harmonic functions, far from decision theory). Its application to decision theory under ambiguity is due to David Schmeidler (1989, "Subjective Probability and Expected Utility Without Additivity," *Econometrica*), with the Möbius-transform/core characterization (connecting Choquet integrals to Γ-minimax) substantially due to Schmeidler's contemporaries and the slightly later Dempster-Shafer theory of evidence (Dempster, 1967; Shafer, 1976) – itself originally developed for statistical inference (Dempster) and artificial intelligence/expert systems (Shafer's popularization), only later recognised as capacity theory's belief-function special case. The Shapley-value connection noted in this section's Remarks (Shapley, 1953) is essentially *contemporaneous* with Choquet's original capacity paper, though the connection between the two 1953 results was not made explicit until decades later – a striking case of simultaneous-but-unconnected discovery.

Theorem: the Möbius transform recovers a capacity's basic probability assignment uniquely

For any capacity ν on a finite Ω, there exists a unique *basic probability assignment* (Möbius transform) $m : 2^\Omega \to \mathbb{R}$ satisfying $m(\emptyset) = 0$, $\sum_{A \subseteq \Omega} m(A) = 1$, and $\nu(B) = \sum_{A \subseteq B} m(A)$ for every $B \subseteq \Omega$; conversely ν is recovered from m by this same summation, and ν is a convex capacity if and only if $m(A) \geq 0$ for every $|A| \geq 2$ (the basic assignment is non-negative on all non-singleton subsets, a condition strictly stronger than ν's own non-negativity).

Proof sketch. The Möbius transform $m(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \nu(B)$ is the standard combinatorial Möbius inversion of the summation $\nu(B) = \sum_{A \subseteq B} m(A)$ over the subset lattice, unique by the general Möbius-inversion theorem for any function on a finite poset; the convexity characterization in terms of non-negative m on non-singletons is a classical result in capacity theory (Shafer, 1976, in the closely related context of belief functions, where m is called the "basic probability assignment" and convex capacities correspond exactly to the belief functions of Dempster–Shafer theory). ■

Example — Computing the Möbius transform for the 3-state capacity of Section 11.3

For the convex capacity of Section 11.3's worked core example ($\nu(\{i\}) = 0.1$ for singletons, $\nu(\{i,j\}) = 0.3$ for pairs, $\nu(\Omega) = 1$): $m(\{i\}) = \nu(\{i\}) = 0.1$ for each singleton; $m(\{i,j\}) = \nu(\{i,j\}) - m(\{i\}) - m(\{j\}) = 0.3 - 0.1 - 0.1 = 0.1$ for each pair; $m(\Omega) = \nu(\Omega) - \sum_{|A|=1} m(A) - \sum_{|A|=2} m(A) = 1 - 3(0.1) - 3(0.1) = 1 - 0.6 = 0.4$. Checking non-negativity on non-singletons: $m(\{i,j\}) = 0.1 \geq 0$ and $m(\Omega) = 0.4 \geq 0$ for all relevant subsets, confirming convexity via the Möbius criterion – an independent confirmation of the direct supermodularity check performed in Section 11.3's original worked example, now via the equivalent (and, for larger $|\Omega|$, often computationally more convenient) Möbius-transform route.

Elicitation implications. The basic probability assignment m has a direct behavioural interpretation in the Dempster–Shafer evidence-theory tradition: $m(A)$ represents the "mass" of belief committed *exactly* to the event A and to no proper subset of A – distinct from $\nu(A)$ itself, which aggregates mass committed to A *or any subset of it*. This distinction matters for elicitation: an expert who reports "I am 40% confident the true state lies somewhere in Ω, beyond what I can attribute to any narrower subset" is directly reporting $m(\Omega) = 0.4$, a more behaviourally primitive quantity than the capacity values $\nu(A)$ themselves, which must then be reconstructed by the summation formula – elicitation protocols building on Dempster–Shafer theory typically elicit m directly for this reason, then derive ν (and hence the core, via Theorem 11.3's equivalence) as a secondary computed quantity.

Historical notes (addendum). The Möbius-transform (basic-probability-assignment) representation of capacities originates in Dempster's (1967) and Shafer's (1976) evidence-theory tradition, developed largely independently of and prior to its adoption in Schmeidler's (1989) decision-theoretic Choquet-expected-utility framework; the cross-fertilization between these two traditions – evidence theory's combinatorial mass-function machinery and decision theory's axiomatic non-additive-probability framework – is a recurring theme noted throughout this Part's historical notes, exemplifying how the same convex-capacity mathematics has been independently motivated and re-derived from rather different starting points (artificial-intelligence-oriented evidence combination on one hand, economic decision theory on the other) before their underlying mathematical unity was fully recognized.

Theorem: the core of a convex capacity equals its representing set of priors

For a convex (supermodular) capacity ν on a finite state space Ω, the *core* $\text{core}(\nu) = \{P \in \Delta(\Omega) : P(A) \geq \nu(A)\ \forall A \subseteq \Omega\}$ is exactly the set Γ of priors such that $\underline{E}_\Gamma[f] = \int f\, d\nu$ for every $f : \Omega \to \mathbb{R}$ – i.e. the Choquet integral with respect to a convex capacity *is* the multiple-priors lower expectation with respect to the capacity's core, making the Choquet (Section 11.3) and Γ-minimax (Section 11.4) frameworks formally identical whenever the capacity is convex.

Proof sketch. This is the Choquet (1953)/Schmeidler (1989) representation result: convexity (supermodularity) of ν is exactly the condition under which the core is nonempty and the Choquet integral coincides with the infimum of ordinary expectations over the core, via the same Möbius-inversion machinery used to compute Choquet integrals as signed sums over the capacity's values on a nested sequence of level sets. ■

Example — Constructing the core explicitly for a 3-state capacity

$\Omega = \{1,2,3\}$, convex capacity $\nu(\emptyset) = 0, \nu(\{1\}) = 0.1, \nu(\{2\}) = 0.1, \nu(\{3\}) = 0.1, \nu(\{1,2\}) = 0.3, \nu(\{1,3\}) = 0.3, \nu(\{2,3\}) = 0.3, \nu(\Omega) = 1$ (one can check supermodularity: $\nu(A \cup B) + \nu(A \cap B) \geq \nu(A) + \nu(B)$ holds for all pairs, e.g. $A = \{1,2\}, B = \{2,3\}$: $\nu(\{1,2,3\}) + \nu(\{2\}) = 1 + 0.1 = 1.1 \geq \nu(\{1,2\}) + \nu(\{2,3\}) = 0.3 + 0.3 = 0.6$ ✓). The core is $\{(p_1,p_2,p_3) : p_i \geq 0.1\ \forall i,\ p_i + p_j \geq 0.3\ \forall i \neq j,\ \sum p_i = 1\}$ – a small triangle strictly inside the probability simplex, symmetric by the capacity's symmetry across states, centred at $(1/3, 1/3, 1/3)$ and shrinking toward that centre as the capacity's "ambiguity gap" ($0.3 - 2 \times 0.1 = 0.1$ for each pair) shrinks toward zero, recovering the degenerate single-prior case $\Gamma = \{(1/3,1/3,1/3)\}$ exactly when the capacity becomes additive ($\nu(A \cup B) = \nu(A) + \nu(B)$ for disjoint A, B).

Practical consequence. The core-equivalence theorem means that, whenever a decision-maker's ambiguity is naturally elicited via a convex capacity (e.g. from a sequence of betting-rate questions, the classical route to eliciting ν experimentally), the resulting Choquet-expected-utility decision rule can be computed *either* via the Möbius/level-set Choquet-integral algorithm of this section's main text *or* via an explicit linear program over the core Γ (minimize $\mathbb{E}_P[f]$ over the core's defining linear inequalities) – the two computational routes are guaranteed to agree exactly by the theorem, giving a practitioner the freedom to choose whichever is computationally more convenient for a given capacity's representation (Möbius coefficients vs. an explicit list of core-defining inequalities).

Historical notes (addendum). The core-equivalence result is due to Schmeidler (1986, "Integral Representation Without Additivity," *Proceedings of the AMS*), building on the cooperative-game-theoretic notion of the "core" introduced decades earlier by Gillies (1953) for entirely different (coalition-payoff-division) purposes – a striking instance of a mathematical structure (the core of a supermodular set function) migrating from cooperative game theory into decision theory under non-additive probability, with the same convexity condition playing analogous stabilizing roles in both contexts (non-emptiness of stable coalition payoff allocations in the game-theoretic origin; non-emptiness of the representing prior set in the decision-theoretic application here).

11.4 Robust Bayesian Decision Theory and Γ-Minimax

Overview

Γ-minimax (robust Bayesian) decision theory posits a *set* Γ of plausible priors (or posteriors) over the state space Θ, and chooses the action minimising the worst-case Bayes risk over Γ. This directly mirrors the ambiguity-set formalism of Part 10, with Γ playing the role of $\mathcal{P}$, but now applied to a decision-maker's *beliefs* rather than to a data-generating process.

Definition: Γ-minimax rule

Given a set of priors Γ over Θ, a loss function $L(a,\theta)$, and an action set $\mathcal{A}$, the Γ-minimax action is

$$a^*_\Gamma = \arg\min_{a\in\mathcal{A}} \sup_{\pi\in\Gamma} \mathbb{E}_{\theta\sim\pi}[L(a,\theta)].$$

If $\Gamma = \{\pi_0\}$ is a singleton, this reduces to standard Bayes risk minimization; if Γ is the set of *all* priors, it reduces to minimax (Section 11.2).

Theorem (Sion's minimax theorem for Γ-minimax). If $\mathcal{A}$ is convex and compact, Γ is convex and weak-$*$ compact (e.g. a moment-constrained set of measures, as in Section 10.1), $L(a,\theta)$ is convex in a and the map $\pi \mapsto \mathbb{E}_\pi[L(a,\theta)]$ is linear (hence continuous) in π for each a, then

$$\min_{a\in\mathcal{A}} \sup_{\pi\in\Gamma} \mathbb{E}_\pi[L(a,\theta)] = \sup_{\pi\in\Gamma} \min_{a\in\mathcal{A}} \mathbb{E}_\pi[L(a,\theta)],$$

i.e. the Γ-minimax action equals the Bayes action against the "worst prior" $\pi^* \in \Gamma$ – the order of minimization and (worst-case) maximization can be exchanged.

Proof sketch. This is a direct application of Sion's generalisation of von Neumann's minimax theorem to infinite-dimensional convex sets, requiring convexity/compactness of both $\mathcal{A}$ and Γ and (quasi-)concavity-convexity of the bilinear-in-the-limit objective $\mathbb{E}_\pi[L(a,\theta)]$ – exactly the conditions verified for the conic ambiguity sets of Section 10.3. ■

Example — Γ-minimax with a contaminated-prior class

Let $\Theta = \{\theta_1,\theta_2\}$, $L(a_1,\theta_1)=0, L(a_1,\theta_2)=10$; $L(a_2,\theta_1)=4, L(a_2,\theta_2)=4$ (a "hedged" action with constant loss). Let $\Gamma = \{(1-\epsilon)\pi_0 + \epsilon Q : Q \text{ arbitrary}\}$ (ϵ-contamination class around a base prior $\pi_0 = (0.7, 0.3)$, $\epsilon = 0.2$). For a_1: $\mathbb{E}_\pi[L(a_1,\theta)] = 10\pi(\theta_2)$. Over Γ, $\pi(\theta_2)$ ranges over $[(1-\epsilon)(0.3), (1-\epsilon)(0.3)+\epsilon] = [0.24, 0.44]$ (since contamination Q can put all its mass on θ_2). So $\sup_{\pi\in\Gamma} \mathbb{E}_\pi[L(a_1,\theta)] = 10\times 0.44 = 4.4$. For a_2: $\mathbb{E}_\pi[L(a_2,\theta)] = 4$ for *every* π (constant loss). Since $4 < 4.4$, the Γ-minimax action is $a^*_\Gamma = a_2$ – the hedged action is preferred once the prior is allowed to be contaminated by up to 20% adversarial mass, even though under the base prior π_0 alone, $\mathbb{E}_{\pi_0}[L(a_1,\theta)] = 10(0.3) = 3 < 4 = \mathbb{E}_{\pi_0}[L(a_2,\theta)]$ would favour a_1.

Computing the worst prior via convex optimization. For Γ a polytope, $\sup_{\pi\in\Gamma} \mathbb{E}_\pi[L(a,\theta)]$ is an LP for fixed a. A cutting-plane algorithm alternates: solve the inner LP for worst prior $\pi^{(k)}$ given $a^{(k)}$, then $a^{(k+1)} = \arg\min_a \mathbb{E}_{\pi^{(k)}}[L(a,\theta)]$ (ordinary Bayes), iterate – the nested-Benders structure that reappears in Section 12.4.

Example — Two-iteration cutting-plane illustration

$\Theta = \{\theta_1, \theta_2\}$, $\pi_0 = (0.7, 0.3)$, $\epsilon = 0.2$, a_1 (losses $0, 10$), a_2 (losses $4, 4$). Iter. 1: $a^{(1)} = a_1$ (Bayes-optimal vs. π_0). Worst prior: $\pi^{(1)} = (0.56, 0.44)$. Iter. 2: $\mathbb{E}_{\pi^{(1)}}[L(a_1)] = 4.4 > 4 = \mathbb{E}_{\pi^{(1)}}[L(a_2)]$, switch to $a^{(2)} = a_2$. Converged in 2 iterations, matching the closed-form $a^* = a_2$.

Connections to robust statistics. The ϵ-contamination class is also the foundational object of robust statistics (Huber, 1964): an estimator's breakdown point is the largest ϵ tolerable under arbitrary contamination. Median: 0.5; mean: 0. Γ-minimax and breakdown-point theory are two applications of the same total-variation ambiguity geometry from Section 10.1.

Connections. The ϵ-contamination class is the decision-theoretic cousin of the ϕ-divergence balls of Section 10.1 (indeed, for the total-variation divergence, ϵ-contamination neighbourhoods and TV-balls of radius ϵ around π_0 coincide). The minimax-theorem-based "worst prior" characterisation gives a constructive recipe – find $\pi^* = \arg\max_{\pi \in \Gamma}$(Bayes risk against π), then play the ordinary Bayes-optimal action against π^* – that operationalises Γ-minimax using standard Bayesian computation, completing the bridge from the purely geometric robustness of Part 9 to fully Bayesian robust decision-making, before Section 11.5 introduces a final, non-probabilistic alternative: info-gap theory.

Theorem: Γ-minimax regret as a hybrid criterion

Define Γ-minimax *regret* as $a^* = \arg\min_a \sup_{\pi \in \Gamma} \left[\max_{a'} \mathbb{E}_\pi[u(a', \theta)] - \mathbb{E}_\pi[u(a, \theta)]\right]$ (a Γ-minimax version of Savage's minimax regret, Section 11.2). Unlike ordinary Γ-minimax (Theorem 11.4), Γ-minimax regret does *not* generally admit a minimax-theorem exchange of $\sup_\pi$ and $\min_a$, because the inner term $\max_{a'} \mathbb{E}_\pi[u(a', \theta)]$ – the "π-best-case benchmark" – itself depends on π in a way that breaks the bilinearity required by Sion's theorem.

Proof sketch. The function $\pi \mapsto \max_{a'} \mathbb{E}_\pi[u(a', \theta)]$ is a *maximum of linear functions of* π, hence *convex* in π (not linear) – it is the support function of the (convex hull of the) set $\{(\mathbb{E}_\pi[u(a', \theta)])_{a'} : \text{varying } a'\}$ evaluated in the "select-the-max" direction. The objective $\sup_\pi$[convex-in-π − linear-in-π] is then a supremum of a convex function, which is itself convex but *not* amenable to the linear-bilinear minimax exchange of Sion's theorem (which requires the saddle function to be convex in one argument and *concave*, not merely any-shape, in the other – here, after the sign flip, the relevant function is convex in π, the *wrong* curvature for the $\sup_\pi$ side of a minimax exchange). ■

Example — A case where the exchange genuinely fails

$\Theta = \{\theta_1, \theta_2\}$, $\Gamma = \{(p, 1-p) : p \in [0, 1]\}$ (all priors, i.e. ordinary minimax regret), actions $a_1 = (10, 0)$, $a_2 = (0, 10)$, $a_3 = (4, 4)$. For each p, $\max_{a'} \mathbb{E}_p[u(a', \theta)] = \max(10p, 10(1-p), 4)$. At $p = 0.5$: $\max(5, 5, 4) = 5$. Regret of a_3 at $p = 0.5$: $5 - 4 = 1$. Regret of a_1 at $p = 0.5$: $5 - 5 = 0$. Regret of a_2 at $p = 0.5$: $5 - 5 = 0$. At $p = 0$: $\max(0, 10, 4) = 10$. Regret of a_3: $10 - 4 = 6$. Regrets of a_1, a_2: $10 - 0 = 10$, $10 - 10 = 0$. The *sup over* $p \in [0, 1]$ of a_3's regret occurs somewhere with regret ≥ 1 (at $p = 0.5$) and potentially higher elsewhere (check $p = 0$: regret 6); a_3's worst-case regret is therefore ≥ 6. Meanwhile a_1's worst-case regret (over p) is $\max(10(1-p) - 10p, \dots)$, evaluating to 10 at $p = 0$ but 0 at $p = 0.5$ – the regret functions are themselves non-trivial (piecewise-linear but with the "kink" from the inner $\max_{a'}$), and direct computation

(rather than a minimax-theorem shortcut) is required to find $\min_a \sup_p$[regret], illustrating the theorem's claim that no easy exchange is available here.

Remarks and extensions. The minimax theorem (Theorem 11.4) requires Γ to be *convex*. Non-convex sets of priors – e.g. a *finite* set $\Gamma = \{\pi_1, \dots, \pi_K\}$ of "scenario" priors representing genuinely distinct hypotheses (rather than a continuum of contamination-type perturbations of a single base prior) – do not admit the minimax-theorem exchange of min and max in general, and Γ-minimax over a finite Γ can require genuinely solving the $\min_a \max_k$ problem directly (e.g. via mixed-integer programming if $\mathcal{A}$ is itself discrete). However, the *convex hull* $\text{conv}(\Gamma) = \{\sum_k \mu_k \pi_k : \mu \in \Delta^{K-1}\}$ *does* satisfy the minimax theorem's hypotheses, and $\sup_{\pi \in \text{conv}(\Gamma)} \mathbb{E}_\pi[L(a,\theta)] = \max_k \mathbb{E}_{\pi_k}[L(a,\theta)]$ (the sup over a simplex of linear functions is attained at a vertex) – so Γ-minimax over $\text{conv}(\Gamma)$ and over the finite Γ itself give the *same* answer (both reduce to $\min_a \max_k \mathbb{E}_{\pi_k}[L(a,\theta)]$), even though only the former technically satisfies the theorem's convexity hypothesis. This reconciliation – finite scenario sets are "secretly convex" for the purposes of Γ-minimax – is a useful practical simplification.

A further remark connects back to Part 10: the Γ-minimax framework with Γ a moment-based ambiguity set (Section 10.1) and $L(a,\theta) = \ell(\mathbf{x} = a, \boldsymbol{\xi} = \theta)$ is *literally* the moment-based DRO problem of Section 10.3 – Γ-minimax decision theory and DRO are, for this choice of Γ, the identical mathematical object viewed through different disciplinary lenses (decision theory vs. optimization), a unification that the closing Section 12.5 will make explicit at the level of the whole framework taxonomy.

Theorem: Γ-minimax with a hierarchical (scenario-indexed) prior set

If $\Gamma = \{\pi(\cdot|\theta_0) : \theta_0 \in \Theta_0\}$ for a (possibly small, finite) hyperparameter set Θ_0, then $\sup_{\pi \in \Gamma} \mathbb{E}_\pi[L(a,\theta)] = \max_{\theta_0 \in \Theta_0} \mathbb{E}_{\pi(\cdot|\theta_0)}[L(a,\theta)]$ – the sup reduces to a maximum over Θ_0, often dramatically cheaper than the LP-based search over an unstructured Γ.

Proof. Immediate from the indexed-union definition of Γ. ■

Example — Scenario-indexed Γ for the hiring-screening case study

Instead of the unstructured ϵ-contamination Γ (requiring the LP-based cutting-plane procedure), suppose the firm identifies 3 named future-pool scenarios ($\theta_0 \in$ {status quo, successful outreach, downturn-driven shift}), each with its own $\pi(\cdot|\theta_0)$. By the theorem, $\sup_{\pi \in \Gamma} \mathbb{E}_\pi[L] = \max(\mathbb{E}_{\pi(\cdot|1)}[L], \mathbb{E}_{\pi(\cdot|2)}[L], \mathbb{E}_{\pi(\cdot|3)}[L])$ – just 3 Bayes-risk evaluations and a max, no LP needed. This named-scenario structure – "stress-test against scenarios 1, 2, 3, plan for the worst" – is, in practice, how many organisations operationalise Γ-minimax, even when the underlying theory (Section 11.4's main text) is framed in terms of more abstract Γ.

Historical notes. The term "Γ-minimax" and its systematic statistical development is associated with Berger (1985, *Statistical Decision Theory and Bayesian Analysis*, a standard graduate textbook), building on a robust-Bayesian research program of the 1980s (Berger, Huber, and others) that sought a middle ground between "pure" Bayesian analysis (a single π_0, criticised as requiring unrealistic precision in prior elicitation) and "pure" frequentist minimax (no prior at all, criticised as overly pessimistic by ignoring genuinely available prior information). The Gilboa-Schmeidler axiomatization (1989, the same year as Schmeidler's Choquet-EU paper – the two papers are often cited together as the twin foundations of modern ambiguity-averse decision theory) provided the formal justification connecting Γ-minimax's pragmatic robust-Bayesian motivation to the Ellsberg-paradox-resolution mo-

tivation of Sections 11.1-11.3, unifying what had been, until 1989, two largely separate literatures (robust Bayesian statistics, and ambiguity-averse decision theory).

Theorem: Γ-minimax regret as Γ shrinks to a point

For a Γ-minimax-regret decision rule $\mathbf{x}^\star_\Gamma = \arg\min_{\mathbf{x}} \sup_{P\in\Gamma} (\mathbb{E}_P[\ell(\mathbf{x}^\star_P,\xi)]-\mathbb{E}_P[\ell(\mathbf{x},\xi)])$ (regret relative to the P-optimal action $\mathbf{x}^\star_P$), as the set of priors Γ shrinks to a singleton $\{P_0\}$ (e.g. via a sequence of nested Γ-sets with diameter $\to 0$ around P_0), $\mathbf{x}^\star_\Gamma \to \mathbf{x}^\star_{P_0}$, the ordinary Bayes-optimal action under P_0 – Γ-minimax regret recovers standard Bayesian decision theory in the no-ambiguity limit, exactly as Γ-minimax (non-regret) expected utility recovers ordinary expected-utility maximization in the same limit.

Proof sketch. As $\Gamma \to \{P_0\}$, the supremum over $P \in \Gamma$ of the regret functional converges (by continuity of the regret functional in P, under mild regularity) to its value at $P = P_0$, namely $\mathbb{E}_{P_0}[\ell(\mathbf{x}^\star_{P_0},\xi)] - \mathbb{E}_{P_0}[\ell(\mathbf{x},\xi)]$, which is minimized over $\mathbf{x}$ exactly at $\mathbf{x} = \mathbf{x}^\star_{P_0}$ (the regret is zero there and strictly positive elsewhere by definition of $\mathbf{x}^\star_{P_0}$ as the unique P_0-optimal action, under standard uniqueness conditions). ■

Example — A finite-Γ illustration of the convergence

Newsvendor-type problem with $\Gamma_r = \{P : \mathbb{E}_P[\xi] \in [50-r, 50+r]\}$ (mean-ambiguity ball of radius r around 50), critical ratio 0.667 (as in Section 12.4's worked VSS example). For $r = 0$, $\Gamma_0 = \{50\}$ exactly, recovering the ordinary stochastic-optimal $x^\star = 66.7$ (the 66.7th percentile under the single distribution with mean 50, here taken as Uniform$(0, 100)$ for concreteness). As r increases from 0, the theorem guarantees only that $x^\star_\Gamma \to 66.7$ as $r \to 0$, *not* that $x^\star_\Gamma$ remains close to 66.7 for every small $r > 0$ – continuity at the limit point does not by itself bound the rate of departure away from it. A separate (and generally harder) sensitivity calculation, differentiating $x^\star_{\Gamma_r}$ with respect to r at $r = 0$, would be needed to assess *how quickly* the Γ-minimax-regret solution moves away from the Bayes solution as ambiguity is introduced – a calculation this section's qualitative continuity theorem does not by itself supply, and a caution against over-interpreting "continuity" as "slowness of change."

Contrast with Γ-minimax (non-regret). It is worth emphasizing that the analogous continuity result for ordinary (non-regret) Γ-minimax expected utility, $\inf_{P\in\Gamma} \mathbb{E}_P[u(\mathbf{x},\theta)]$, converges to $\mathbb{E}_{P_0}[u(\mathbf{x},\theta)]$ under the same shrinking-Γ limit by an essentially identical argument – both Γ-minimax variants (regret and non-regret) share this reassuring continuity property, even though they can differ substantially from each other (and from ordinary Bayes) for non-trivial Γ, as already noted in this section's main-text discussion of regret-based vs. value-based robust Bayesian criteria.

Historical notes (addendum). The shrinking-Γ continuity property, while intuitively expected and straightforward to prove, is an important sanity check for any robust decision-theoretic framework – a framework failing this property would imply that the very act of acknowledging (even infinitesimally small) prior ambiguity could produce a discontinuous jump in recommended action, a pathology that would undermine practical trust in the Γ-minimax apparatus. Berger's (1985, *Statistical Decision Theory and Bayesian Analysis*) systematic treatment of robust Bayesian methods explicitly verifies this continuity property as a basic desideratum for any proposed robustness criterion, a standard this section's Γ-minimax-regret framework satisfies as shown above.

11.5 Info-Gap Decision Theory

Overview

Info-gap decision theory (Ben-Haim, 2001) dispenses with probability altogether. Uncertainty is modelled by a nested family of sets $\mathcal{U}(\alpha, \mathbf{u}_0)$ (the "horizon of uncertainty" $\alpha \geq 0$ around a best estimate $\mathbf{u}_0$), and robustness is the *largest* α for which a performance requirement is guaranteed – an explicitly non-probabilistic, set-based formalism that closes the loop back to Part 9's uncertainty sets.

Definition: Robustness and opportuneness functions

Given a model $u(\mathbf{x}, \mathbf{a})$ (e.g. performance of decision $\mathbf{x}$ under true parameter $\mathbf{a}$), best estimate $\mathbf{u}_0$, nested uncertainty sets $\mathcal{U}(\alpha, \mathbf{u}_0) \subseteq \mathcal{U}(\alpha', \mathbf{u}_0)$ for $\alpha \leq \alpha'$ (with $\mathcal{U}(0, \mathbf{u}_0) = \{\mathbf{u}_0\}$), and a critical performance threshold r_c, the *robustness function* is

$$\hat{\alpha}(\mathbf{x}, r_c) = \max\left\{\alpha \geq 0 : \min_{\mathbf{a} \in \mathcal{U}(\alpha, \mathbf{u}_0)} u(\mathbf{x}, \mathbf{a}) \geq r_c\right\},$$

the largest uncertainty horizon over which $\mathbf{x}$ is guaranteed to meet the requirement $u \geq r_c$.

Theorem (info-gap robustness as an inverse support function). For $\mathcal{U}(\alpha, \mathbf{u}_0) = \{\mathbf{a} : \|\mathbf{a} - \mathbf{u}_0\| \leq \alpha\}$ (nested balls) and $u(\mathbf{x}, \mathbf{a}) = \mathbf{c}(\mathbf{x})^\top \mathbf{a}$ linear in $\mathbf{a}$ (cf. Section 9.1), $\min_{\mathbf{a} \in \mathcal{U}(\alpha, \mathbf{u}_0)} u(\mathbf{x}, \mathbf{a}) = \mathbf{c}(\mathbf{x})^\top \mathbf{u}_0 - \alpha \|\mathbf{c}(\mathbf{x})\|_*$ (dual norm), so

$$\hat{\alpha}(\mathbf{x}, r_c) = \frac{\mathbf{c}(\mathbf{x})^\top \mathbf{u}_0 - r_c}{\|\mathbf{c}(\mathbf{x})\|_*} \quad \text{(if positive, else } \hat{\alpha} = 0),$$

an explicit closed form – the robustness is the "slack" $\mathbf{c}(\mathbf{x})^\top \mathbf{u}_0 - r_c$ divided by the "sensitivity" $\|\mathbf{c}(\mathbf{x})\|_*$, exactly the support-function calculus of Section 9.1 read in reverse (solving for the radius α rather than for the worst-case value).

Proof. $\min_{\|\mathbf{a}-\mathbf{u}_0\| \leq \alpha} \mathbf{c}^\top \mathbf{a} = \mathbf{c}^\top \mathbf{u}_0 + \min_{\|\delta\| \leq \alpha} \mathbf{c}^\top \delta = \mathbf{c}^\top \mathbf{u}_0 - \alpha \|\mathbf{c}\|_*$ by the same Hölder/Cauchy–Schwarz argument as Eq. (9.1.ell)/(10.5). Setting this $\geq r_c$ and solving for the largest such α gives the stated formula. ■

Example — Comparing two designs by info-gap robustness

$\mathbf{u}_0 = (10, 5)^\top$ (best-estimate parameters), $r_c = 40$ (minimum required performance), norm $= \ell_2$ (so $\|\cdot\|_* = \ell_2$ too). Design $\mathbf{x}_1$ has $\mathbf{c}(\mathbf{x}_1) = (3, 2)^\top$: $\mathbf{c}^\top \mathbf{u}_0 = 30 + 10 = 40$. But $\mathbf{c}^\top \mathbf{u}_0 - r_c = 0$, so $\hat{\alpha}(\mathbf{x}_1, r_c) = 0$ – design $\mathbf{x}_1$ exactly meets the requirement at the best estimate with *zero* robustness to any deviation. Design $\mathbf{x}_2$ has $\mathbf{c}(\mathbf{x}_2) = (2, 2)^\top$: $\mathbf{c}^\top \mathbf{u}_0 = 20 + 10 = 30 < r_c = 40$ – design $\mathbf{x}_2$ does not even meet the requirement at the best estimate, so $\hat{\alpha}(\mathbf{x}_2, r_c) = 0$ as well (the formula would give a negative numerator, truncated to 0). Design $\mathbf{x}_3$ has $\mathbf{c}(\mathbf{x}_3) = (5, 3)^\top$: $\mathbf{c}^\top \mathbf{u}_0 = 50 + 15 = 65$, $\|\mathbf{c}(\mathbf{x}_3)\|_2 = \sqrt{34} \approx 5.83$, $\hat{\alpha}(\mathbf{x}_3, r_c) = (65 - 40)/5.83 \approx 4.29$. Despite $\mathbf{x}_1$ achieving the requirement *exactly* at the nominal estimate (often perceived as "efficient"), it has zero robustness; $\mathbf{x}_3$, which appears to "over-deliver" at the nominal estimate, can tolerate parameter deviations up to $\alpha \approx 4.29$ – info-gap theory formalises the intuition that nominal-optimality and robustness are different objectives, often in tension (the "robustness–performance trade-off" that closes Part 11).

Trade-off curves and the opportuneness function. Dual to robustness $\hat{\alpha}$ is opportuneness $\hat{\beta}(\mathbf{x}, r_w) = \min\{\alpha : \max_{\mathbf{a}\in\mathcal{U}(\alpha,\mathbf{u}_0)} u(\mathbf{x},\mathbf{a}) \geq r_w\}$ for a windfall target r_w: the smallest horizon at which the windfall *becomes possible*. Plotting $\hat{\alpha}$ vs. $\hat{\beta}$ across designs traces a robustness-opportuneness frontier.

Example — Opportuneness for the three designs

With $r_w = 80$, $\hat{\beta}(\mathbf{x}, r_w) = \frac{r_w - \mathbf{c}^\top \mathbf{u}_0}{\|\mathbf{c}\|_*}$. For $\mathbf{x}_3$: $(80-65)/5.83 \approx 2.57$. For $\mathbf{x}_1$: $(80-40)/3.606 \approx 11.09$. So $\mathbf{x}_3$ has both higher robustness (≈ 4.29 vs. 0) and better opportuneness (2.57 vs. 11.09) – $\mathbf{x}_3$ dominates $\mathbf{x}_1$ on both axes, possible because $\mathbf{x}_1$ was a poorly-chosen baseline with no slack.

Info-gap vs. probability: a reconciliation. The closed-form robustness function of Theorem 11.5 is identical in form to the support function of Section 9.1 and the SOC reformulation of Section 12.3 – α plays the role of the ellipsoid radius / safety factor. The interpretation differs, but the mathematics – support functions and dual norms – is shared, an observation Section 12.5 elevates to an organizing table for the whole volume.

Synthesis of Part 11. Sections 11.1–11.5 trace a progression from a purely qualitative violation of SEU (Ellsberg), through simple worst-case/regret criteria (Wald, Savage, Hurwicz), to fully axiomatized non-additive-measure (Choquet) and set-of-priors (Γ-minimax) frameworks, ending with the explicitly non-probabilistic, set-based info-gap formalism that reconnects directly to the geometric language of Part 9. Each framework can be seen as choosing a different "ambiguity primitive" – a capacity, a set of priors, or a nested family of uncertainty sets – on which to ground decisions when P itself is unknown, setting the stage for Part 12's treatment of *multi-period* decisions under uncertainty, where these single-shot frameworks must be composed sequentially with recourse.

Part 12

Scenario Analysis and Stochastic Programming

Multi-scenario formal frameworks for sequential decision-making under uncertainty, with recourse.

Theorem: info-gap robustness under a general (non-ball) nested family

If $\mathcal{U}(\alpha, \mathbf{u}_0) = \{\mathbf{u}_0 + \alpha\mathbf{d} : \mathbf{d} \in \mathcal{D}\}$ for a fixed convex set $\mathcal{D}$ (not necessarily a norm ball) with $\mathbf{0} \in \mathcal{D}$, and $u(\mathbf{x}, \mathbf{a}) = \mathbf{c}(\mathbf{x})^\top \mathbf{a}$, then

$$\hat{\alpha}(\mathbf{x}, r_c) = \frac{\mathbf{c}(\mathbf{x})^\top \mathbf{u}_0 - r_c}{-\inf_{\mathbf{d}\in\mathcal{D}} \mathbf{c}(\mathbf{x})^\top \mathbf{d}} = \frac{\mathbf{c}(\mathbf{x})^\top \mathbf{u}_0 - r_c}{\sigma_{-\mathcal{D}}(\mathbf{c}(\mathbf{x}))}$$

(if the numerator and the support function are both positive; 0 otherwise) – the ball-based formula of Theorem 11.5 is the special case $\mathcal{D} = \{\|\mathbf{d}\| \leq 1\}$, $\sigma_{-\mathcal{D}} = \|\cdot\|_*$.

Proof. $\min_{\mathbf{a}\in\mathcal{U}(\alpha,\mathbf{u}_0)} \mathbf{c}^\top\mathbf{a} = \mathbf{c}^\top\mathbf{u}_0 + \alpha \min_{\mathbf{d}\in\mathcal{D}} \mathbf{c}^\top\mathbf{d} = \mathbf{c}^\top\mathbf{u}_0 - \alpha\sigma_{-\mathcal{D}}(\mathbf{c})$ (using $\min_{\mathbf{d}\in\mathcal{D}} \mathbf{c}^\top\mathbf{d} = -\sup_{\mathbf{d}\in\mathcal{D}}(-\mathbf{c})^\top\mathbf{d} = -\sigma_{\mathcal{D}}(-\mathbf{c}) = -\sigma_{-\mathcal{D}}(\mathbf{c})$ by definition of support function and a sign convention). Setting this $\geq r_c$ and solving for α gives the stated formula. ■

Example — Asymmetric uncertainty: a one-sided info-gap model

Suppose load uncertainty is known to be *one-sided* (loads can only *increase* beyond the nominal estimate, e.g. due to ratchet-type effects in traffic growth): $\mathcal{D} = \{\mathbf{d} \geq \mathbf{0}\} \cap \{\|\mathbf{d}\|_2 \leq 1\}$ (a quarter-disc, for $\mathbb{R}^2$). For $\mathbf{c}(\mathbf{x}) = (3,2)^\top$ (both positive), $\sigma_{-\mathcal{D}}(\mathbf{c}) = \sup_{\mathbf{d}\in\mathcal{D}}(-\mathbf{c})^\top\mathbf{d} = \sup_{\mathbf{d}\geq 0,\|\mathbf{d}\|\leq 1}(-3d_1 - 2d_2) = 0$ (attained at $\mathbf{d} = \mathbf{0}$, since $-3d_1 - 2d_2 \leq 0$ for $\mathbf{d} \geq \mathbf{0}$, with equality only at $\mathbf{0}$). So $\sigma_{-\mathcal{D}}(\mathbf{c}) = 0$, giving $\hat{\alpha} = \frac{\mathbf{c}^\top\mathbf{u}_0 - r_c}{0} = +\infty$ whenever $\mathbf{c}^\top\mathbf{u}_0 > r_c$ – **infinite robustness**: if uncertainty can only *increase* $\mathbf{a}$ (hence increase $\mathbf{c}^\top\mathbf{a}$ for $\mathbf{c} \geq \mathbf{0}$) and the constraint is $\mathbf{c}^\top\mathbf{a} \geq r_c$ (capacity exceeds requirement), *no* amount of (one-sided, adverse-only-in-the-other-direction) uncertainty can violate it. This degenerate-but-instructive case shows that the *direction* of uncertainty relative to the constraint's sense matters enormously – a one-sided uncertainty model that happens to point "the helpful way" relative to a given constraint yields trivial (infinite) robustness, a modelling artifact to be aware of when $\mathcal{D}$ is asymmetric.

Remarks and extensions. Info-gap theory has been the subject of methodological debate, particularly regarding its claimed independence from probability theory. Critics (e.g. Sniedovich) have argued that, as the closed-form result of Theorem 11.5 illustrates, info-gap robustness functions are often mathematically equivalent to specific (if implicit) probabilistic or set-based robust-optimization quantities – so that info-gap's practical recommendations could, in principle, always be derived from the RO/DRO frameworks of Parts 9–10 with an appropriate (if sometimes non-obvious) choice of uncertainty set or ambiguity set. Proponents respond that info-gap's *operational* value lies precisely in *not requiring* the decision-maker to commit, even implicitly, to such a choice – the nested family $\mathcal{U}(\alpha, \mathbf{u}_0)$ for *all* $\alpha \geq 0$ is specified once, and $\hat{\alpha}$ is computed without needing to pre-select a "right" α (analogous to how a p-value summarizes evidence across a range of significance thresholds without committing to one in advance).

This debate is, in microcosm, the debate underlying the entire taxonomy of Section 12.5: every framework in this volume can be seen as a particular *operationalisation* of the same underlying mathematical structures (support functions, dual norms, conic duality), differing mainly in *which* quantities are treated as "given" (a fixed $\mathcal{U}$? a fixed β? a fixed Γ? or, as in info-gap, none of these, with the answer expressed as a *function* of the not-yet-fixed quantity α). Recognising this shared structure – rather than treating each framework as a separate "tool" to be selected from a toolbox – is the central methodological message of this volume, made fully explicit in the unifying table of Section 12.5.

Theorem: info-gap robustness curves are piecewise-linear and non-increasing in r_c

$\hat{\alpha}(\mathbf{x}, r_c)$ is non-increasing in r_c, and for the linear-u/ball-$\mathcal{U}$ case of Theorem 11.5 is piecewise linear with slope $-1/\|\mathbf{c}(\mathbf{x})\|_*$ wherever $\hat{\alpha} > 0$, becoming flat at 0 once $\mathbf{c}(\mathbf{x})^\top\mathbf{u}_0 \leq r_c$.

Proof. From $\hat{\alpha} = \frac{\mathbf{c}(\mathbf{x})^\top\mathbf{u}_0 - r_c}{\|\mathbf{c}(\mathbf{x})\|_*}$, $\frac{\partial\hat{\alpha}}{\partial r_c} = -\frac{1}{\|\mathbf{c}(\mathbf{x})\|_*} < 0$; for $r_c \geq \mathbf{c}(\mathbf{x})^\top\mathbf{u}_0$, $\hat{\alpha} = 0$ by the implicit non-negativity truncation. ∎

Example — Reading a robustness curve as a design-comparison menu

For $\mathbf{x}_3$ ($\mathbf{c}^\top\mathbf{u}_0 = 65, \|\mathbf{c}\|_2 \approx 5.83$): $\hat{\alpha}(\mathbf{x}_3, r_c) = (65 - r_c)/5.83$ for $r_c \leq 65$, 0 otherwise. At $r_c = 40, 50, 60, 65$: $\hat{\alpha} \approx 4.29, 2.57, 0.86, 0$. This curve is a *menu* – for any possible future regulatory requirement r_c, robustness is read off directly without re-solving. Compar-

ing such piecewise-linear menus across designs (each with its own slope $-1/\|\mathbf{c}(\mathbf{x})\|_*$ and x-intercept $\mathbf{c}(\mathbf{x})^\top \mathbf{u}_0$) gives a direct visual comparison tool: *flatter* curves (larger $\|\mathbf{c}(\mathbf{x})\|_*$) degrade more slowly as r_c rises, preferable when future requirement increases are anticipated.

Historical notes. Info-gap decision theory was developed by Yakov Ben-Haim starting in the early 1990s and presented comprehensively in his 2001 book *Information-Gap Decision Theory: Decisions Under Severe Uncertainty*. Unlike most frameworks in this Part (which emerged from economics/statistics), info-gap originated in *engineering* (Ben-Haim's background is in mechanical/structural engineering and applied mathematics), reflecting engineering's practical need for robustness criteria that do not require probability distributions for quantities (material properties, future loads, model parameters) that are often genuinely unknown in a way that resists even the "set of priors" formalization of Γ-minimax. The Sniedovich critique mentioned in this section's Remarks (Sniedovich, 2007 onward, a sustained methodological debate spanning multiple journal exchanges) represents one of the more visible foundational controversies in the broader robust-decision-making literature, and the debate's resolution (or lack thereof) remains, two decades on, more a matter of methodological/philosophical preference than settled mathematical fact – an unusually open-ended note on which to close a section, but an honest reflection of the field's current state.

Theorem: the robustness-opportuneness trade-off curve is monotone

For an info-gap robustness function $\hat{\alpha}(\mathbf{x}, r_c) = \max\{\alpha : \min_{u \in \mathcal{U}(\alpha,\xi_0)} \text{performance}(\mathbf{x}, u) \geq r_c\}$ and the corresponding opportuneness function $\hat{\beta}(\mathbf{x}, r_w) = \min\{\alpha : \max_{u \in \mathcal{U}(\alpha,\xi_0)} \text{performance}(\mathbf{x}, u) \geq r_w\}$ (the "best-case" counterpart of robustness), for a fixed action $\mathbf{x}$, $\hat{\alpha}(\mathbf{x}, \cdot)$ is non-increasing in the required critical performance level r_c, while $\hat{\beta}(\mathbf{x}, \cdot)$ is non-decreasing in the desired windfall level r_w – robustness erodes (more uncertainty must be tolerated to feel "safe enough") as the safety bar r_c is raised, while opportuneness similarly recedes (more uncertainty is needed before a windfall becomes *possible*) as the windfall bar r_w is raised.

Proof. Both monotonicity claims follow directly from the nested structure of the info-gap uncertainty sets $\mathcal{U}(\alpha, \xi_0)$ (increasing in α, Section 11.5's main-text definition) combined with the fact that a higher required performance level r_c can only be guaranteed (the min over $\mathcal{U}(\alpha, \xi_0)$ exceeding r_c) for a *smaller* feasible set of α (smaller uncertainty horizons), hence $\hat{\alpha}$ is non-increasing in r_c; symmetrically, a higher desired windfall r_w requires the *max* over $\mathcal{U}(\alpha, \xi_0)$ to reach further, which (since $\mathcal{U}$ is nested and growing in α) requires *larger*, not smaller, α, giving non-decreasing $\hat{\beta}$ in r_w. ■

Example — Plotting both curves for a single design choice

For a structural design $\mathbf{x}$ with nominal load capacity $\xi_0 = 100$ and a multiplicative info-gap model $\mathcal{U}(\alpha, \xi_0) = [\xi_0(1-\alpha), \xi_0(1+\alpha)]$ for the true (unknown) load demand, with performance defined as capacity minus demand (survival requires demand $\leq$ capacity, so the relevant worst case is demand *exceeding* the nominal value ξ_0, an upper-tail event): with the design's true capacity at 120, robustness to a critical-survival requirement (demand must not exceed 120) corresponds to $\hat{\alpha}$ solving $\xi_0(1+\hat{\alpha}) = 120 \implies 100(1+\hat{\alpha}) = 120 \implies \hat{\alpha} = 0.2$ (can tolerate up to 20% demand uncertainty before failing this survival bar). For opportuneness with wind-

fall bar $r_w = 70$ (capacity exceeding demand by a comfortable margin, i.e. demand as low as 70): $\hat{\beta}$ solves $100(1-\hat{\beta}) = 70 \implies \hat{\beta} = 0.3$ – a windfall this favourable *requires* at least 30% downward demand uncertainty to even be possible, illustrating the theorem's claimed monotonicity (here, in the reverse/best-case direction) and giving the design's decision-maker a complete two-sided uncertainty picture: robust against 20% adverse swings, but no realistic prospect of the favourable 30% windfall scenario without an equally large (and, ex ante, equally uncertain) downward demand swing.

The combined robustness-opportuneness report. Practically deployed info-gap analyses (Ben-Haim's monograph emphasizes this point repeatedly) report *both* curves side by side for competing design alternatives, since a design with high robustness but negligible opportuneness (a "play it safe, no upside" profile) may be preferred or dis-preferred relative to a design with lower robustness but substantial opportuneness (a "calculated bet" profile), depending on the decision-maker's risk appetite – info-gap theory, in this dual-curve form, supplies the descriptive apparatus for this trade-off without requiring (per its founding motivation, Section 11.5's main text) a probability distribution over the uncertain quantity at all.

Historical notes (addendum). The opportuneness function, while less emphasized in applied info-gap case studies than the robustness function, is present from the framework's earliest systematic treatment (Ben-Haim, 1st edition 2001) as the necessary "other half" of a complete uncertainty picture; its relative neglect in practice mirrors a broader asymmetry, also visible in the robust-optimization literature of Part 9, where worst-case (downside) analysis receives substantially more attention than best-case (upside) analysis, despite the two being formally dual constructions in every framework surveyed in this volume.

12.1 Two-Stage Stochastic Programming with Recourse

Overview

Two-stage stochastic programming with recourse is the probabilistic counterpart of the adjustable robust optimization of Section 9.5: a first-stage decision $\mathbf{x}$ is made before $\boldsymbol{\xi}$ is observed, and a second-stage recourse decision $\mathbf{y}(\boldsymbol{\xi})$ minimises expected (rather than worst-case) cost given the realisation of $\boldsymbol{\xi}$.

Definition: Two-stage stochastic LP with recourse

$$\min_{\mathbf{x}\geq 0} \ \mathbf{c}^\top \mathbf{x} + \mathbb{E}_{\boldsymbol{\xi}}[Q(\mathbf{x},\boldsymbol{\xi})] \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b},$$

where the *recourse function* $Q(\mathbf{x},\boldsymbol{\xi})$ is the value of the second-stage problem

$$Q(\mathbf{x},\boldsymbol{\xi}) = \min_{\mathbf{y}\geq 0} \ \mathbf{q}(\boldsymbol{\xi})^\top \mathbf{y} \quad \text{s.t.} \quad \mathbf{W}(\boldsymbol{\xi})\mathbf{y} = \mathbf{h}(\boldsymbol{\xi}) - \mathbf{T}(\boldsymbol{\xi})\mathbf{x}.$$

$\mathbf{x}$ is chosen "here and now"; $\mathbf{y}$ is chosen "wait and see," after $\boldsymbol{\xi}$ is revealed.

Theorem (convexity of Q and the expected recourse problem). If $\mathbf{W}(\boldsymbol{\xi}) \equiv \mathbf{W}$ is fixed (fixed recourse) and the feasible region of the second stage is nonempty for all relevant $(\mathbf{x},\boldsymbol{\xi})$ (*relatively complete recourse*), then $Q(\mathbf{x},\boldsymbol{\xi})$ is convex and piecewise linear in $\mathbf{x}$ for each $\boldsymbol{\xi}$, and $\mathbb{E}_{\boldsymbol{\xi}}[Q(\mathbf{x},\boldsymbol{\xi})]$ is convex in $\mathbf{x}$, so the overall two-stage problem is a convex program (an LP if $\boldsymbol{\xi}$ has finite support).

Proof sketch. For fixed $\boldsymbol{\xi}$, $Q(\mathbf{x}, \boldsymbol{\xi})$ is the optimal value of an LP whose right-hand side $\mathbf{h}(\boldsymbol{\xi}) - \mathbf{T}(\boldsymbol{\xi})\mathbf{x}$ is affine in $\mathbf{x}$; the optimal value of an LP is a convex (in fact piecewise-linear, by LP value-function theory) function of the right-hand side, hence convex in $\mathbf{x}$ as a composition of a convex function with an affine map. Convexity is preserved under (finite or integrable) expectation. ■

Example — The newsvendor problem as a two-stage stochastic LP

Order quantity $x \geq 0$ at unit cost c; demand ξ is random; unsold units incur holding cost h per unit, shortage incurs penalty p per unit. Second stage: $Q(x, \xi) = \min_{y_1, y_2 \geq 0}\{hy_1 + py_2 : y_1 \geq x - \xi,\ y_2 \geq \xi - x\} = h\max(0, x-\xi) + p\max(0, \xi - x)$. The problem $\min_x cx + \mathbb{E}_\xi[Q(x,\xi)]$ has first-order condition (for continuous ξ with CDF F): $c + hF(x) - p(1 - F(x)) = 0 \Rightarrow F(x^*) = \frac{p-c}{p+h}$, the classical *critical ratio* (newsvendor) formula. E.g. $c = 2, h = 1, p = 5$: $F(x^*) = (5-2)/(5+1) = 0.5$, so x^* is the *median* of demand – if $\xi \sim \text{Uniform}(0, 100)$, $x^* = 50$.

Multi-item and capacity-constrained extensions. The scalar newsvendor extends to n items with $\sum_j x_j \leq C$. With independent ξ_j and $C = \infty$ the problem separates (Theorem 12.1 item-by-item). The capacity constraint couples items via shadow price $\lambda \geq 0$: $F_j(x_j^*) = \frac{p_j - c_j - \lambda}{p_j + h_j}$, with λ chosen so $\sum_j x_j^*(\lambda) = C$.

Example — Two-item capacitated newsvendor

Items A $(c, h, p = 2, 1, 5$, ratio $0.5)$ and B $(1, 1, 4$, ratio $0.6)$, both $\sim \text{Uniform}(0, 100)$. Unconstrained: $x_A^* = 50$, $x_B^* = 60$. With $C = 90 < 110$: $x_A(\lambda) = 50 - 16.67\lambda$, $x_B(\lambda) = 60 - 20\lambda$. Setting sum $= 90$: $\lambda \approx 0.545$, $x_A^* \approx 40.9$, $x_B^* \approx 49.1$. Item B (higher ratio) retains a larger share of constrained capacity.

Risk-averse two-stage programming. Replacing $\mathbb{E}[Q]$ by $\text{CVaR}_\beta[Q]$ preserves convexity and LP-representability (Rockafellar-Uryasev), so Theorem 12.1 extends unchanged. CVaR is the tightest convex upper bound on the chance-constraint indicator at level β, linking risk-averse SP to Section 12.3's chance constraints.

Connections. The two-stage structure here is formally identical to the adjustable RO of Section 9.5, with the inner $\max_{\mathbf{a} \in \mathcal{U}}$ (worst-case) replaced by $\mathbb{E}_{\boldsymbol{\xi}}$ (expectation). The same affine-decision-rule approximation idea applies (approximate $\mathbf{y}(\boldsymbol{\xi})$ by an affine function for tractability when $\boldsymbol{\xi}$ has continuous/high-dimensional support), but for finite-support $\boldsymbol{\xi}$ (the *scenario* representation developed in Section 12.2) the problem is an exact, finite LP – the "extensive form" – whose size grows linearly in the number of scenarios, motivating scenario reduction techniques.

Theorem: value of the stochastic solution (VSS)

Define z_{EV} as the cost of the *expected-value* solution (solve the deterministic problem with $\boldsymbol{\xi} = \mathbb{E}[\boldsymbol{\xi}]$, then evaluate its *expected* recourse cost under the true distribution of $\boldsymbol{\xi}$), and z_{SP} as the true two-stage stochastic-programming optimum (Theorem 12.1). Then $\text{VSS} := z_{\text{EV}} - z_{\text{SP}} \geq 0$ (by optimality of z_{SP} over all $\mathbf{x}$, including the EV solution's $\mathbf{x}$), with equality iff the EV solution is itself SP-optimal (e.g. if $Q(\mathbf{x}, \boldsymbol{\xi})$ is affine in $\boldsymbol{\xi}$ for the relevant $\mathbf{x}$, so $\mathbb{E}[Q(\mathbf{x}, \boldsymbol{\xi})] = Q(\mathbf{x}, \mathbb{E}[\boldsymbol{\xi}])$, "certainty equivalence").

Proof. $z_{\text{SP}} = \min_{\mathbf{x}}[\mathbf{c}^\top \mathbf{x} + \mathbb{E}_{\boldsymbol{\xi}}[Q(\mathbf{x}, \boldsymbol{\xi})]] \leq \mathbf{c}^\top \mathbf{x}_{\text{EV}} + \mathbb{E}_{\boldsymbol{\xi}}[Q(\mathbf{x}_{\text{EV}}, \boldsymbol{\xi})] = z_{\text{EV}}$ since $\mathbf{x}_{\text{EV}}$ is a (suboptimal, in general) candidate for the SP minimization. ■

Example — Computing VSS for the newsvendor

For $c = 2, h = 1, p = 5$, $\xi \sim \text{Uniform}(0, 100)$ (so $\mathbb{E}[\xi] = 50$). EV solution: $x_{\text{EV}} = \mathbb{E}[\xi] = 50$ (order the mean demand – a common naive heuristic). But the SP-optimal is $x^*_{\text{SP}} = 50$ *too* (computed in Section 12.1: the critical ratio is 0.5, and the median of Uniform$(0, 100)$ is also 50)! So in this particular instance, VSS $= 0$ – the EV solution happens to coincide with the SP-optimal, because for Uniform$(0, 100)$, mean $=$ median $=$ the 50th-percentile critical-ratio target (a coincidence of this specific critical ratio 0.5 and this specific symmetric distribution). For a *different* critical ratio, say $p = 8, h = 1, c = 2$ (ratio $(8-2)/(8+1) = 6/9 \approx 0.667$): $x^*_{\text{SP}} = 66.7$ (the 66.7th percentile of Uniform(0,100)) $\neq x_{\text{EV}} = 50$. Using $\mathbb{E}[Q(x, \xi)] = \frac{hx^2 + p(100-x)^2}{200}$ for Uniform(0,100): $\mathbb{E}[Q(50, \xi)] = \frac{1(2500)+8(2500)}{200} = \frac{2500+20000}{200} = 112.5$, and $\mathbb{E}[Q(66.7, \xi)] = \frac{1(4448.9)+8(1108.9)}{200} \approx \frac{4448.9+8871.1}{200} = 66.6$. Total cost: $z_{\text{EV}} = cx_{\text{EV}} + \mathbb{E}[Q(50, \xi)] = 2(50) + 112.5 = 212.5$; $z_{\text{SP}} = cx_{\text{SP}} + \mathbb{E}[Q(66.7, \xi)] \approx 2(66.7) + 66.6 = 200.0$. VSS $= 212.5 - 200.0 = 12.5 > 0$ ✓, as the theorem guarantees – ordering the mean demand ($x_{\text{EV}} = 50$) costs \$12.5 more (in expectation) than the SP-optimal $x^*_{\text{SP}} \approx 66.7$, the quantitative value of properly accounting for the asymmetric cost structure ($p/h = 8$) rather than naively ordering to the mean.

Remarks and extensions. The newsvendor problem, despite its simplicity, has served throughout this volume's literature as the canonical "Drosophila" of stochastic and robust optimization: Section 9.4's price-of-robustness, Section 10.2's WDRO regularization, Section 11.5's info-gap robustness, and Section 12.5's robust-vs-stochastic comparison are all, in this textbook and in the broader literature, frequently illustrated on newsvendor-type problems precisely because the critical-ratio formula provides an easily-computable "ground truth" against which each framework's robust/ambiguous/info-gap variant can be transparently compared. Readers are encouraged to revisit each of these sections' worked examples and verify that they are, in fact, all variations on the single newsvendor cost structure $h \max(0, x - \xi) + p \max(0, \xi - x)$ introduced here – an exercise in recognising the shared structure underlying superficially different "examples."

A further remark on relatively complete recourse (assumed in Theorem 12.1): if the second-stage problem can be *infeasible* for some $(\mathbf{x}, \boldsymbol{\xi})$ (e.g. if $\mathbf{x}$ is too small to satisfy a hard minimum-service constraint regardless of recourse), the recourse function $Q(\mathbf{x}, \boldsymbol{\xi}) = +\infty$ for such $(\mathbf{x}, \boldsymbol{\xi})$, and $\mathbb{E}_{\boldsymbol{\xi}}[Q(\mathbf{x}, \boldsymbol{\xi})] = +\infty$ unless Pr[infeasible] $= 0$ – effectively imposing a hard (probability-zero) feasibility requirement on $\mathbf{x}$, which can itself be reformulated as a chance constraint at level $\beta = 0$ (Section 12.3's $\beta \to 0$ limit, previewed here).

Theorem: the newsvendor recourse is a scaled call-option payoff

$Q(x, \xi) = h \max(0, x - \xi) + p \max(0, \xi - x) = h(x - \xi) + (h + p) \max(0, \xi - x)$. Hence $\mathbb{E}_{\boldsymbol{\xi}}[Q(x, \boldsymbol{\xi})] = h(x - \mathbb{E}[\xi]) + (h + p)C(x)$ where $C(x) = \mathbb{E}[\max(0, \xi - x)]$ is exactly the expected payoff of a European call option with strike x – the newsvendor problem and call-option pricing share an identical mathematical core.

Proof. Using $\max(0, x - \xi) = (x - \xi) + \max(0, \xi - x)$: $h \max(0, x - \xi) + p \max(0, \xi - x) = h[(x - \xi) + \max(0, \xi - x)] + p \max(0, \xi - x) = h(x - \xi) + (h + p) \max(0, \xi - x)$. ■

Example — Critical ratio via option-pricing calculus

Using $C'(x) = F(x) - 1$ (the call's "delta with respect to strike"), differentiating $cx + h(x - \mathbb{E}[\xi]) + (h+p)C(x)$ w.r.t. x and setting to zero: $c + h + (h+p)(F(x^*) - 1) = 0 \iff F(x^*) = \frac{p-c}{p+h}$ ✓, recovering Theorem 12.1's critical ratio via option-pricing "Greeks" – a second derivation route useful for practitioners with a finance background.

Historical notes. Two-stage stochastic programming with recourse traces to Dantzig (1955, "Linear Programming Under Uncertainty"), making it – alongside Wald's maximin (1945, Section 11.2's notes) – among the very earliest formal frameworks for optimization under uncertainty, predating both the Ellsberg paradox (1961) and modern robust optimization (1970s onward) by years to decades. The newsvendor problem itself is far older still as an *informal* problem (variants appear in inventory-theory writings from the early 20th century and arguably earlier in informal merchant practice), but its formalization as a two-stage stochastic LP with the critical-ratio solution is generally credited to the Dantzig-era operations research community (1950s-60s) that also produced the simplex method and the broader edifice of linear programming – two-stage SP with recourse is, in this sense, almost as old as linear programming itself, having developed in close historical parallel rather than as a later "extension."

Theorem: convexity and subdifferentiability of the expected recourse function

For the two-stage SP recourse function $Q(\mathbf{x}, \xi) = \min_{\mathbf{y} \geq 0}\{\mathbf{q}^\top \mathbf{y} : \mathbf{W}\mathbf{y} = \xi - \mathbf{T}\mathbf{x}\}$ (linear recourse, fixed recourse matrix $\mathbf{W}$), $Q(\cdot, \xi)$ is convex and piecewise-linear in $\mathbf{x}$ for each fixed ξ, and the expected recourse $\mathcal{Q}(\mathbf{x}) = \mathbb{E}_\xi[Q(\mathbf{x}, \xi)]$ is convex in $\mathbf{x}$ with subdifferential $\partial\mathcal{Q}(\mathbf{x}) = \mathbb{E}_\xi[-\mathbf{T}^\top \boldsymbol{\pi}^\star(\mathbf{x}, \xi)]$, where $\boldsymbol{\pi}^\star(\mathbf{x}, \xi)$ is an optimal dual vector for the recourse LP at $(\mathbf{x}, \xi)$.

Proof sketch. $Q(\cdot, \xi)$ is the value function of a parametric LP with the parameter $(\xi - \mathbf{T}\mathbf{x})$ entering only the right-hand side affinely; such value functions are well known to be convex and piecewise-linear in the right-hand-side parameter (a standard parametric-LP sensitivity result), hence convex in $\mathbf{x}$ since $\mathbf{x} \mapsto \xi - \mathbf{T}\mathbf{x}$ is affine and convexity is preserved under affine pre-composition. The subdifferential formula follows from Danskin's theorem (or LP sensitivity analysis directly): the dual optimal $\boldsymbol{\pi}^\star$ gives the local sensitivity of the optimal value to right-hand-side perturbations, here transmitted through $-\mathbf{T}^\top$ by the chain rule, and expectation preserves convexity (an expectation of convex functions is convex) and, under standard regularity (e.g. dominated convergence), commutes with the subdifferential. ■

Example — Computing the subgradient in the newsvendor case

For the scalar newsvendor recourse $Q(x, \xi) = h\max(x - \xi, 0) + p\max(\xi - x, 0)$ (Section 12.1's main text), the optimal dual/subgradient with respect to x is h if $x > \xi$ (overage active) and $-p$ if $x < \xi$ (underage active); hence $\partial_x \mathbb{E}_\xi[Q(x, \xi)] = h\Pr[\xi < x] - p\Pr[\xi > x] = hF(x) - p(1 - F(x)) = (h+p)F(x) - p$. Setting this subgradient to zero recovers exactly the critical-ratio first-order condition $F(x^\star) = p/(h+p)$ derived directly in Section 12.1's main text – the general convex-analysis machinery of this theorem, applied to the scalar newsvendor special case, reproduces the elementary calculus-based derivation exactly, confirming the theorem's subdifferential formula concretely and illustrating how the general multi-dimensional SP sensitivity theory specializes cleanly to the textbook scalar case.

Algorithmic significance. The subdifferential formula is precisely what makes Benders

decomposition (used throughout Section 12.4's SDDP algorithm) computationally viable: each "cut" added to the value-function approximation $\underline{V}_t$ is constructed directly from a dual vector $\boldsymbol{\pi}^\star$ obtained while solving a recourse subproblem, exploiting exactly this theorem's identification of $\partial \mathcal{Q}(\mathbf{x})$ with an expectation of recourse-LP dual vectors – the entire SDDP machinery of Section 12.4 is, at bottom, an algorithmic exploitation of this section's convexity-and-subdifferentiability theorem, applied stage by stage.

Historical notes (addendum). The value-function convexity and subdifferential characterization for two-stage linear recourse problems is classical, tracing to Wets (1974, "Stochastic Programs with Fixed Recourse") and Rockafellar–Wets's systematic development through the 1970s-80s, predating by over a decade the SDDP algorithm (Pereira–Pinto, 1991, Section 12.4's historical notes) that would later exploit it computationally at industrial scale – another instance, as in the Cramér/large-deviations connection noted in Section 9.4's addendum, of foundational convex-analytic results substantially preceding their large-scale algorithmic exploitation.

12.2 Scenario Generation and Reduction Techniques

Overview

When $\boldsymbol{\xi}$ has continuous or very large discrete support, the expectation $\mathbb{E}_{\boldsymbol{\xi}}[Q(\mathbf{x}, \boldsymbol{\xi})]$ must be approximated by a finite set of *scenarios* $\{(\boldsymbol{\xi}^{(s)}, p_s)\}_{s=1}^{S}$ with $\sum_s p_s = 1$. This section covers principled methods for generating such scenario sets and for *reducing* an initially large set while controlling the approximation error.

Scenario generation methods. (i) *Monte Carlo sampling*: draw $\boldsymbol{\xi}^{(s)} \sim P$, set $p_s = 1/S$; error $O(S^{-1/2})$ by the CLT. (ii) *Moment matching*: choose $\{\boldsymbol{\xi}^{(s)}, p_s\}$ to exactly match the first k moments of P with $S \ll$ (number of moment constraints would naively require), via numerical optimization. (iii) *Quasi-Monte Carlo / low-discrepancy sequences*: deterministic point sets with $O(S^{-1}(\log S)^d)$ discrepancy, faster convergence for smooth Q.

Definition: Scenario reduction via Wasserstein distance

Given an initial scenario set $\{(\boldsymbol{\xi}^{(s)}, p_s)\}_{s=1}^{S}$ (defining empirical measure $\hat{P}_S$), *scenario reduction* seeks a smaller set $\{(\boldsymbol{\xi}^{(s')}, p_{s'})\}_{s'=1}^{S'}$ ($S' < S$, typically a subset of the original scenarios with redistributed probabilities) defining $\hat{P}_{S'}$, minimising $W_1(\hat{P}_S, \hat{P}_{S'})$ – the same Wasserstein distance of Section 10.2, now used to measure scenario-set approximation error rather than distributional ambiguity.

Theorem (optimal reduction via the transportation problem). For fixed retained scenario indices $R \subseteq \{1, \dots, S\}$ ($|R| = S'$), the optimal redistribution of probabilities $p_{s'}$ for $s' \in R$ and the resulting $W_1(\hat{P}_S, \hat{P}_{S'})$ are given by solving the optimal transport (Kantorovich) problem between $\hat{P}_S$ (fixed) and measures supported on $\{\boldsymbol{\xi}^{(s')}\}_{s' \in R}$; the optimal redistribution assigns each deleted scenario's probability mass to its *nearest* (in the cost metric d) retained scenario. Selecting R itself to minimize W_1 is combinatorial (choosing S' of S points), but a greedy "delete the scenario causing smallest W_1 increase" heuristic has provable approximation guarantees for several metric structures.

Example — Reducing 5 scenarios to 2

Scenarios $\boldsymbol{\xi}^{(s)} \in \{1,2,3,4,5\}$ (1-D for simplicity) each with $p_s = 0.2$. Suppose we retain $R = \{1,5\}$ (the extremes). The nearest-retained-point redistribution sends $\xi = 2 \to 1$ (distance 1) and $\xi = 3,4 \to 5$ (distances 2,1) – but $\xi = 3$ is equidistant from 1 and 5 (distance 2 each); send it to either, say to 5. Resulting $\hat{P}_{S'}$: $p_1 = 0.2 + 0.2 = 0.4$ (absorbing $\xi = 2$), $p_5 = 0.2 + 0.2 + 0.2 = 0.6$ (absorbing $\xi = 3,4$). $W_1(\hat{P}_S, \hat{P}_{S'}) = \sum_s p_s \cdot d(\xi^{(s)}$, its assigned retained point$) = 0.2(0) + 0.2(1) + 0.2(2) + 0.2(1) + 0.2(0) = 0.2(0+1+2+1+0) = 0.8$. Compare retaining $R = \{2,4\}$ (more central): $\xi = 1 \to 2$ (dist 1), $\xi = 3 \to 2$ or 4 (dist 1 either way, say split or assign to 2), $\xi = 5 \to 4$ (dist 1); all distances are ≤ 1, giving $W_1 \leq 0.2 \times (1+0+1+0+1) = 0.6 < 0.8$ – the centrally-located retained set achieves a strictly smaller approximation error, illustrating that "extremes-only" reduction (a common naive heuristic) is not generally optimal.

Scenario generation for dependent random variables. For correlated $\boldsymbol{\xi}$, copula-based generation samples the dependence structure separately from the marginals via the probability integral transform, preserving both correct marginals and dependence at no extra cost over independent sampling.

Example — Correlated two-asset scenario generation

Two assets $\xi_1, \xi_2 \sim \mathcal{N}(0,1)$, Gaussian copula $\rho = 0.7$. With $z_1 = 1, z_1^{\perp} = 0.5$: $z_2 = 0.7(1) + \sqrt{0.51}(0.5) \approx 1.057$. For Gaussian marginals $(\xi_1, \xi_2) = (1, 1.057)$ directly. For $\xi_2 \sim \text{Exp}(1)$: $u_2 = \Phi(1.057) \approx 0.8546$, $\xi_2 = -\log(1-0.8546) \approx 1.929$ – $\rho = 0.7$ governs rank correlation regardless of marginal shape.

Out-of-sample validation of reduced scenario sets. A reduced set should be validated by checking $\mathbf{x}^*(\hat{P}_{S'})$'s expected cost under the *original* $\hat{P}_S$ against $\mathbf{x}^*(\hat{P}_S)$'s – the exact, problem-specific value of information lost to reduction, complementing the Wasserstein-distance bound which can be loose.

Discussion. Scenario reduction is the computational bridge between the abstract two-stage formulation of Section 12.1 (defined for general P) and the solvable finite LP "extensive form." The Wasserstein-distance criterion connects directly to the WDRO machinery of Section 10.2: a solution $\mathbf{x}^*$ that is optimal for the reduced problem $\hat{P}_{S'}$ has expected-cost performance under the true P that differs from the $\hat{P}_{S'}$-optimal value by at most a Lipschitz-constant-times-$W_1(\hat{P}_{S'}, P)$ term – so scenario-reduction error and DRO-style robustness bounds are two sides of the same Wasserstein-continuity coin.

Theorem: scenario reduction error bounds the optimality gap

If $\mathbf{x}^{S'}$ is optimal for the reduced problem (using $\hat{P}_{S'}$) and $Q(\mathbf{x}, \boldsymbol{\xi})$ is L_Q-Lipschitz in $\boldsymbol{\xi}$ uniformly over $\mathbf{x}$ in the feasible region, then

$$\mathbb{E}_{\hat{P}_S}[Q(\mathbf{x}^{S'}, \boldsymbol{\xi})] - \mathbb{E}_{\hat{P}_S}[Q(\mathbf{x}^{S}, \boldsymbol{\xi})] \leq 2L_Q \, W_1(\hat{P}_S, \hat{P}_{S'}),$$

where $\mathbf{x}^S$ is the (unreduced) optimal solution – the optimality gap from using the reduced scenario set is bounded by twice the Lipschitz constant times the Wasserstein reduction error.

Proof sketch. By Theorem 10.2-type Lipschitz/Wasserstein bounds applied twice: $|\mathbb{E}_{\hat{P}_S}[Q(\mathbf{x}, \boldsymbol{\xi})] - \mathbb{E}_{\hat{P}_{S'}}[Q(\mathbf{x}, \boldsymbol{\xi})]| \leq L_Q W_1(\hat{P}_S, \hat{P}_{S'})$ for *any fixed* $\mathbf{x}$ (in particular for $\mathbf{x}^{S'}$ and $\mathbf{x}^S$). Then $\mathbb{E}_{\hat{P}_S}[Q(\mathbf{x}^{S'}, \cdot)] \leq$

$\mathbb{E}_{\hat{P}_{S'}}[Q(\mathbf{x}^{S'},\cdot)] + L_Q W_1 \le \mathbb{E}_{\hat{P}_{S'}}[Q(\mathbf{x}^{S},\cdot)] + L_Q W_1$ (optimality of $\mathbf{x}^{S'}$ for $\hat{P}_{S'}$) $\le \mathbb{E}_{\hat{P}_S}[Q(\mathbf{x}^{S},\cdot)] + 2L_Q W_1$ (Lipschitz bound again). ■

Example — Using the bound to certify a reduction's quality

If $Q(\mathbf{x},\xi) = h\max(0, x-\xi) + p\max(0,\xi-x)$ (newsvendor recourse) with $h=1, p=5$, then Q is Lipschitz in ξ with constant $L_Q = \max(h,p) = 5$ (the steeper of the two linear pieces). For the Section 12.2 reduction example ($S=5 \to S'=2$, $W_1 \approx 0.6$ for the better choice $R=\{2,4\}$), the bound gives an optimality gap $\le 2 \times 5 \times 0.6 = 6$ (in whatever cost units Q is measured). If the SP-optimal cost itself is, say, ≈ 112 (as in the previous section's VSS example, taking the corrected/intended value), a gap bound of 6 represents $\approx 5\%$ – a quantifiable, *a priori* (computable before solving the reduced problem) certificate of the reduction's quality, directly addressing the "how good is good enough" question left informal in Section 12.2's main discussion of validation.

Remarks and extensions. Scenario reduction, as presented in this section, operates on a *fixed* initial scenario set. An alternative, increasingly common in large-scale applications, is *scenario clustering*: rather than selecting a subset of an initial (large) set, cluster all N original samples (e.g. via k-means on $\boldsymbol{\xi}$-space, possibly weighted by an estimate of each sample's "influence" on $Q(\mathbf{x},\cdot)$) into $S' \ll N$ clusters, and use cluster centroids (with probabilities equal to cluster sizes $/N$) as the reduced scenario set. Clustering-based reduction can be seen as an approximate, computationally cheap surrogate for the exact Wasserstein-optimal reduction of Theorem 12.2 (cluster centroids approximately minimise within-cluster transport cost, the discrete analogue of k-means' minimisation of within-cluster variance), and scales to N in the millions where exact transportation-based reduction would be computationally prohibitive.

A second remark concerns the interaction between scenario reduction and the multi-stage setting of Section 12.4: reducing the scenario *tree* (rather than a flat scenario set) must respect the tree's nested structure – reducing branches at stage t affects the conditional scenario sets available at all stages $> t$. Tree-reduction algorithms typically proceed recursively from the leaves inward, applying flat (Section 12.2-style) reduction at each node's set of children, a natural but non-trivial generalisation that is an active topic in the multi-stage stochastic programming literature.

Theorem: centroid-based reduction and Jensen's inequality

If $Q(\mathbf{x},\boldsymbol{\xi})$ is convex in $\boldsymbol{\xi}$ and reduction maps each deleted scenario to the *conditional-expectation* (centroid) of its assigned cluster, then $\mathbb{E}_{\hat{P}_S}[Q(\mathbf{x},\boldsymbol{\xi})] \ge \mathbb{E}_{\hat{P}_{S'}}[Q(\mathbf{x},\boldsymbol{\xi}')]$ – centroid-based reduction *systematically underestimates* the true expected recourse cost, a *signed* bias (unlike the unsigned Lipschitz bound).

Proof sketch. By Jensen's inequality for convex $Q(\mathbf{x},\cdot)$: $\mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})|T(\boldsymbol{\xi})] \ge Q(\mathbf{x},\mathbb{E}[\boldsymbol{\xi}|T(\boldsymbol{\xi})])$; taking outer expectations gives $\mathbb{E}_{\hat{P}_S}[Q] \ge \mathbb{E}_{\hat{P}_{S'}}[Q]$ when $\hat{P}_{S'}$'s scenarios are these conditional means. ■

Example — Centroid- vs. nearest-point reduction, revisited

For $R=\{1,5\}$ with $\xi=2$ assigned to cluster-1 and $\xi=3,4$ to cluster-5: cluster-1's centroid is $\mathbb{E}[\xi|\xi\in\{1,2\}] = 1.5$ (vs. nearest-point's 1), cluster-5's centroid is $\mathbb{E}[\xi|\xi\in\{3,4,5\}] = 4$ (vs. 5). By the theorem, this centroid-based reduced problem *underestimates* $\mathbb{E}_{\hat{P}_S}[Q(\mathbf{x},\xi)]$ for convex Q – a known-direction bias a practitioner can

use for a corrective adjustment, rather than relying solely on the unsigned Lipschitz bound.

Historical notes. Scenario generation and reduction as a distinct methodological topic (rather than an ad-hoc preprocessing step) emerged in the 1990s-2000s stochastic programming literature, with Wasserstein-distance-based reduction specifically developed by Dupačová, Gröwe-Kuska, and Römisch (2003, "Scenario Reduction in Stochastic Programming"). This 2003 work *predates* the 2017-2018 Wasserstein-DRO "rediscovery" wave noted in Section 10.2's historical notes by over a decade – the stochastic programming community had been using Wasserstein (then often called Kantorovich) distances for scenario-set comparison well before the machine-learning community's adversarial-robustness-motivated rediscovery of essentially the same distance for ambiguity sets, a clear instance of the same mathematical tool being independently developed within two adjacent but, at the time, poorly-cross-referenced sub-disciplines of optimization.

Theorem: moment-matching scenario generation as a constrained optimization problem

The moment-matching approach to scenario generation – constructing a small scenario set $\{\xi_i, p_i\}_{i=1}^{N'}$ such that its empirical moments up to order K exactly match a set of target moments $\{m_1, \dots, m_K\}$ (typically estimated from historical data or a fitted parametric model) – can be formulated as the feasibility problem

$$\text{find } \{\xi_i, p_i\}_{i=1}^{N'} : \sum_i p_i\, \xi_i^k = m_k,\ k = 1, \dots, K, \quad \sum_i p_i = 1,\ p_i \geq 0,$$

which, for fixed scenario *locations* $\{\xi_i\}$ and unknown probabilities $\{p_i\}$ only, reduces to a linear feasibility problem (hence solvable by LP, or by a moment-constrained entropy-maximization problem if a unique, most-non-committal solution among the feasible set is desired).

Proof. With ξ_i fixed, each moment-matching constraint $\sum_i p_i \xi_i^k = m_k$ is linear in the unknowns p_i; together with the simplex constraints $\sum p_i = 1, p_i \geq 0$, the full system is a linear feasibility problem, directly amenable to LP feasibility-checking (or, with an entropy objective $-\sum p_i \ln p_i$ added, a convex but non-linear program with a unique maximum-entropy solution whenever the feasible set is nonempty). ■

Example — A four-scenario moment-matching construction

To match a target mean $m_1 = 50$, variance $m_2 - m_1^2 = 200$ (so $m_2 = 2700$), and zero skewness ($m_3 = m_1^3 + 3m_1(m_2 - m_1^2) \cdot 0 + 0 =$ consistent with symmetry) using $N' = 4$ scenarios at fixed, symmetric locations $\xi \in \{50 - 2\sigma, 50 - \sigma, 50 + \sigma, 50 + 2\sigma\}$ with $\sigma = \sqrt{200} \approx 14.14$: by symmetry, set $p_1 = p_4 = q$, $p_2 = p_3 = r$ with $2q + 2r = 1$. Matching the variance: $2q(2\sigma)^2 + 2r(\sigma)^2 = 200 \implies 8q\sigma^2 + 2r\sigma^2 = 200/$(after dividing by $\sigma^2 = 200$) $\implies 8q + 2r = 1$. Combined with $2q + 2r = 1 \implies r = 0.5 - q$: substituting, $8q + 2(0.5 - q) = 1 \implies 8q + 1 - 2q = 1 \implies 6q = 0 \implies q = 0$, $r = 0.5$ – i.e. the moment-matching constraints for this particular four-point symmetric layout collapse onto only the two inner points $\xi = 50 \pm \sigma$, each with probability 0.5, illustrating that not every superficially reasonable scenario-location choice yields a non-degenerate solution using all the allotted points: the practitioner must either solve for both locations *and* probabilities jointly (a substantially harder, generally non-convex problem) or choose locations more carefully relative to the target moments before fixing them.

Joint location-and-probability optimization. When both $\{\xi_i\}$ and $\{p_i\}$ are free, moment-matching scenario generation becomes the harder problem of finding *any* discrete distribution on N' points matching K moments – a classical problem in the theory of Gaussian quadrature and moment problems (closely related to the Isii/Chebyshev-type extremal moment problems of Section 10.3's addendum and Section 12.3's historical notes): a N'-point discrete distribution has $2N'-1$ free parameters (locations and probabilities, minus one for the simplex constraint), so matching $K \leq 2N'-1$ moments is generically (though not always, as the degenerate example above shows) achievable, with N' scenarios able to match up to $2N'-1$ moments – the discrete-distribution analogue of Gaussian quadrature's classical result that N' quadrature points can exactly integrate polynomials up to degree $2N'-1$.

Connection to Part 10's moment-DRO. This moment-matching scenario-generation technique and Part 10's moment-based ambiguity sets (Section 10.1, 10.3) are, in a precise sense, dual constructions: scenario generation *fixes* the moments and searches for a (small, discrete) representing distribution, while moment-DRO *fixes* a reference distribution's moments and considers the entire *set* of distributions sharing them as an ambiguity set to optimize against – the same underlying moment-problem mathematics (generalized Chebyshev-Markov bounds, SDP-representable moment cones) underlies both constructions, a connection made explicit in Volume V's treatment of the moment problem's role across stochastic and robust optimization alike.

Historical notes (addendum). Moment-matching scenario generation for stochastic programming was systematized by Høyland and Wallace (2001, "Generating Scenario Trees for Multistage Decision Problems"), whose nonlinear-programming-based joint location-and-probability solver remains in wide practical use despite the generic non-convexity of the joint problem noted above; subsequent work (Høyland, Kaut, Wallace, 2003) developed specialized heuristics to mitigate the non-convexity, though a fully reliable globally-optimal algorithm for the general joint problem remains, like several other open questions catalogued in this volume's Part 13, an active area of methodological development.

Theorem: error bound for scenario reduction via the Kantorovich (Wasserstein) distance

If a scenario set $\{\xi_i, p_i\}_{i=1}^{N}$ is reduced to a smaller set $\{\tilde{\xi}_j, \tilde{p}_j\}_{j=1}^{N'}$ $(N' < N)$ such that the Wasserstein distance $W_1(\hat{P}_N, \hat{P}_{N'}) \leq \delta$, then for any L-Lipschitz recourse function $Q(\mathbf{x}, \cdot)$ (Lipschitz in ξ, uniformly in $\mathbf{x}$ over the feasible region), the resulting expected-recourse approximation error is bounded: $|\mathbb{E}_{\hat{P}_N}[Q(\mathbf{x},\xi)] - \mathbb{E}_{\hat{P}_{N'}}[Q(\mathbf{x},\xi)]| \leq L\delta$ for every feasible $\mathbf{x}$.

Proof. This is a direct application of the Kantorovich-duality characterization of W_1 (Appendix B of this volume): $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq L \cdot W_1(P, Q)$ for any L-Lipschitz f, applied with $f = Q(\mathbf{x}, \cdot)$, $P = \hat{P}_N$, $Q = \hat{P}_{N'}$. ■

Example — Selecting a reduction tolerance from a Lipschitz estimate

For the newsvendor recourse $Q(x, \xi) = h\max(x-\xi, 0) + p\max(\xi - x, 0)$, the Lipschitz constant in ξ (uniformly over x) is $L = \max(h, p)$ (the recourse cost changes by at most $\max(h,p)$ per unit change in ξ, attained whichever regime –overage or underage– is active). For $h = 1, p = 8$ (Section 12.1's asymmetric worked example), $L = 8$; a scenario-reduction procedure targeting an expected-recourse approximation error of at most \$0.50 should therefore choose a reduced scenario set with $W_1(\hat{P}_N, \hat{P}_{N'}) \leq 0.5/8 = 0.0625$ (in demand units) – giving a concrete, theorem-backed stopping criterion for the greedy/heuristic scenario-reduction algorithms surveyed in

this section's main text, rather than the more common practice of fixing N' heuristically (e.g. "keep the top 20 scenarios by probability mass") without a quantified approximation-error guarantee.

Trade-off with multi-stage tree size. Combining this bound with Section 12.4's tree-size analysis: for a T-stage scenario tree with branching factor b at each stage, naively reducing each stage's branching from b to $b' < b$ (to control the b^{T-1} blow-up) introduces a per-stage Wasserstein error δ_t; assuming (for a rough composability estimate) that per-stage errors propagate roughly additively along the value-function recursion (an approximation, not an exact result, since the true error propagation depends on the recourse function's stage-wise Lipschitz structure), a T-stage tree reduced this way at every stage incurs a cumulative error bound of order $L \sum_{t=1}^{T} \delta_t$ – making explicit the tension between this section's scenario-reduction error control and Section 12.4's exponential tree-size pressure: aggressive branching reduction (small b', hence large savings in b'^{T-1}) typically increases each δ_t, and the practitioner must balance the two effects rather than treating them as independent design choices.

Historical notes (addendum). The Wasserstein-distance-based error bound for scenario reduction, while a natural application of Kantorovich duality, was formalized specifically for scenario-tree reduction by Heitsch and Römisch (2003, "Scenario Reduction Algorithms in Stochastic Programming"), whose forward/backward greedy reduction heuristics (cited in this section's main text) are explicitly designed to greedily minimize the Wasserstein distance between the original and reduced scenario sets – making the theorem above not merely a post-hoc error bound but the actual design criterion underlying the most widely used scenario-reduction algorithms in practice, a direct (if not always advertised) point of contact between the classical scenario-reduction literature and the Wasserstein-DRO machinery of Part 10.

12.3 Chance-Constrained Programming

Overview

Chance-constrained programming (CCP) requires constraints to hold with at least a specified probability $1-\beta$ rather than almost surely (as in robust optimization, Part 9) or in expectation (as in stochastic programming with recourse, Sections 12.1–12.2): $\Pr[\mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x} \le b] \ge 1-\beta$. This section develops tractable reformulations for Gaussian and distributionally-robust chance constraints.

Theorem: Gaussian chance constraint as an SOC constraint

If $\mathbf{a}(\boldsymbol{\xi}) = \bar{\mathbf{a}} + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \square)$, then

$$\Pr\left[(\bar{\mathbf{a}} + \boldsymbol{\xi})^\top \mathbf{x} \le b\right] \ge 1-\beta \iff \bar{\mathbf{a}}^\top \mathbf{x} + \Phi^{-1}(1-\beta)\left\|\square^{1/2}\mathbf{x}\right\|_2 \le b,$$

where Φ is the standard normal CDF – an SOC constraint, of identical algebraic form to the ellipsoidal robust counterpart of Eq. (9.1.ell), with the ellipsoid radius $\left\|\mathbf{P}^\top \mathbf{x}\right\|_2 \to \Phi^{-1}(1-\beta)\left\|\square^{1/2}\mathbf{x}\right\|_2$.

Proof. $(\bar{\mathbf{a}} + \boldsymbol{\xi})^\top \mathbf{x} = \bar{\mathbf{a}}^\top \mathbf{x} + \boldsymbol{\xi}^\top \mathbf{x}$, and $\boldsymbol{\xi}^\top \mathbf{x} \sim \mathcal{N}(0, \mathbf{x}^\top \square \mathbf{x})$, so $\boldsymbol{\xi}^\top \mathbf{x} = \left\|\square^{1/2}\mathbf{x}\right\|_2 Z$ for $Z \sim \mathcal{N}(0,1)$. The constraint becomes $\Pr[\bar{\mathbf{a}}^\top \mathbf{x} + \left\|\square^{1/2}\mathbf{x}\right\|_2 Z \le b] \ge 1-\beta \iff \Pr[Z \le \frac{b - \bar{\mathbf{a}}^\top \mathbf{x}}{\left\|\square^{1/2}\mathbf{x}\right\|_2}] \ge 1-\beta \iff$

$\frac{b-\bar{\mathbf{a}}^\top \mathbf{x}}{\|\square^{1/2}\mathbf{x}\|_2} \geq \Phi^{-1}(1-\beta)$, rearranging to the stated SOC form (assuming $\|\square^{1/2}\mathbf{x}\|_2 > 0$). ■

Distribution-free bounds via Chebyshev/Cantelli. Without normality, the one-sided Chebyshev (Cantelli) inequality gives a conservative but distribution-free reformulation: $\Pr[\boldsymbol{\xi}^\top \mathbf{x} \geq t] \leq \frac{\sigma_{\mathbf{x}}^2}{\sigma_{\mathbf{x}}^2+t^2}$ for $t > 0$, $\sigma_{\mathbf{x}}^2 = \mathbf{x}^\top \square \mathbf{x}$, leading to the sufficient condition $\bar{\mathbf{a}}^\top \mathbf{x} + \sqrt{\frac{1-\beta}{\beta}} \|\square^{1/2}\mathbf{x}\|_2 \leq b$ – the same SOC *form* as the Gaussian case, but with $\Phi^{-1}(1-\beta)$ replaced by the larger, distribution-free factor $\sqrt{(1-\beta)/\beta}$.

Example — Comparing Gaussian and Chebyshev safety factors

For $\beta = 0.05$ (95% reliability): Gaussian factor $\Phi^{-1}(0.95) \approx 1.645$. Chebyshev factor $\sqrt{(1-0.05)/0.05} = \sqrt{19} \approx 4.359$ – about $2.65\times$ larger. For $\beta = 0.01$ (99%): Gaussian $\Phi^{-1}(0.99) \approx 2.326$; Chebyshev $\sqrt{99} \approx 9.95$, about $4.3\times$ larger. The Chebyshev bound's conservatism grows rapidly as $\beta \to 0$ (the price of dropping the normality assumption), illustrating the value of even partial distributional information – a recurring theme connecting back to the moment-based ambiguity sets of Section 10.1, of which the Chebyshev-based CCP is precisely the chance-constrained analogue.

Sample-average approximation of chance constraints. With only samples $\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_N$, the chance constraint can be approximated by requiring it on $\geq \lceil (1-\beta)N \rceil$ samples – combinatorial in general, but the SOC reformulations of this section serve as tractable relaxations, with SAA an asymptotically exact benchmark for their conservatism.

Example — SAA vs. Gaussian SOC on non-Gaussian data

Suppose $\boldsymbol{\xi}^\top \mathbf{x} \sim \text{Exp}(1)$. For $\beta = 0.05$: true requirement $t \geq -\log(0.05) \approx 2.996$. Gaussian-SOC (mean 1, var 1): $t \geq 1 + 1.645 = 2.645$ – *under-protective*. Chebyshev-SOC: $t \geq 1 + \sqrt{19} \approx 5.359$ – over-protective but safe. Gaussian risks under-protection on skewed data; Chebyshev is always safe but loose.

DRO chance constraints. Combining ambiguity sets with chance constraints: $\inf_{P \in \mathcal{P}} \Pr_P[\ldots] \geq 1-\beta$. For moment-based $\mathcal{P}$, the uniform Chebyshev bound *is* the moment-based DRO chance constraint – unifying Gaussian-exact and Chebyshev-conservative as the $\mathcal{P} = \{P^*\}$ and $\mathcal{P} = \{P : \text{cov} \preceq \square\}$ extremes.

Joint chance constraints and Bonferroni. For *joint* constraints $\Pr[\bigcap_{i=1}^m \{\mathbf{a}_i^\top \mathbf{x} \leq b_i\}] \geq 1-\beta$, the Bonferroni bound $\Pr[\bigcup_i \{\text{violation}_i\}] \leq \sum_i \Pr[\text{violation}_i]$ gives the sufficient (but generally conservative) decomposition into m individual chance constraints each at level β/m (or any β_i with $\sum_i \beta_i = \beta$), each reformulable via the single-constraint results above – reducing the joint CCP to m independent SOC constraints, at the cost of (typically mild) additional conservatism, the chance-constrained analogue of the per-row decomposition of the robust counterpart in Section 9.2.

Theorem: chance constraints and Value-at-Risk

The chance constraint $\Pr[\mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x} \leq b] \geq 1-\beta$ is equivalent to $\text{VaR}_\beta(\mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x}) \leq b$, where $\text{VaR}_\beta(Z) = \inf\{t : \Pr[Z \leq t] \geq 1-\beta\}$ is the Value-at-Risk at level β – i.e. chance-constrained programming is, equivalently, *VaR-constrained* programming, connecting Section 12.3 directly to the risk-measure literature (and to CVaR_β, mentioned in Section 12.1's remarks, which satisfies $\text{CVaR}_\beta(Z) \geq \text{VaR}_\beta(Z)$ always, with equality only in special cases – so a CVaR-constraint $\text{CVaR}_\beta(\cdot) \leq b$ is *more conservative* than the corresponding chance/VaR constraint).

Proof. By definition, $\mathrm{VaR}_\beta(Z) \le b \iff \inf\{t : \Pr[Z \le t] \ge 1-\beta\} \le b \iff \Pr[Z \le b] \ge 1-\beta$ (using right-continuity of the CDF: if the infimum is $\le b$, then $\Pr[Z \le b] \ge \Pr[Z \le \mathrm{VaR}_\beta(Z)] \ge 1-\beta$, and conversely). ■

Example — CVaR-constraint as a conservative surrogate

For $Z \sim \mathcal{N}(0,1)$, $\beta = 0.05$: $\mathrm{VaR}_{0.05}(Z) = \Phi^{-1}(0.95) \approx 1.645$. $\mathrm{CVaR}_{0.05}(Z) = \mathbb{E}[Z \mid Z \ge 1.645]$, which for a standard normal equals $\frac{\phi(1.645)}{0.05} \approx \frac{0.103}{0.05} \approx 2.06$ (using the standard normal density ϕ). So a $\mathrm{CVaR}_{0.05}(\cdot) \le b$ constraint requires $\bar{\mathbf{a}}^\top \mathbf{x} + 2.06 \left\| \square^{1/2} \mathbf{x} \right\|_2 \le b$ – a *larger* safety factor (2.06) than the chance/VaR constraint's 1.645 (Theorem 12.3), but still far smaller than the Chebyshev factor $\sqrt{19} \approx 4.36$ for the same β. CVaR thus occupies a middle ground – more conservative than exact-Gaussian VaR (accounting for tail risk *beyond* the β-quantile, which VaR ignores entirely) but less conservative than the fully distribution-free Chebyshev bound – and, being a coherent risk measure (subadditive, unlike VaR), CVaR-constraints compose better across multiple risks, a further reason for CVaR's popularity in financial risk management beyond its LP-representability (Section 12.1's remark).

Remarks and extensions. The Gaussian chance-constraint SOC reformulation (Theorem 12.3) extends to *joint* Gaussian chance constraints $\Pr[\bigcap_i \{\cdot\}] \ge 1-\beta$ without the (sometimes loose) Bonferroni decomposition, *if* the constraint violations are themselves jointly Gaussian – the joint probability can then be computed via the multivariate normal CDF $\Phi_m(\cdot; \mathbf{R})$ (correlation matrix $\mathbf{R}$), and the resulting reformulation, while no longer a simple SOC, can be handled via the SOC-representable *approximations* to Φ_m^{-1} developed in the chance-constrained optimization literature (e.g. via the Bonferroni bound's refinements, or via copula-based decompositions analogous to Section 12.2's scenario-generation copulas).

A second remark concerns *ambiguous* chance constraints where even β itself is uncertain (e.g. a regulatory requirement specifies "$\beta \in [0.01, 0.05]$, exact value to be determined"). Requiring the chance constraint to hold for the *worst* β in this range is straightforward (use $\beta = 0.01$, the most stringent); but if the decision-maker additionally has a *distribution* over plausible β values (reflecting, e.g., the probability that a stricter regulatory standard will be adopted), a Γ-minimax-style (Section 11.4) treatment of β-ambiguity becomes relevant – yet another instance of this volume's recurring theme that "ambiguity about ambiguity" can itself be handled by the same toolkit, one level up.

Theorem: chance-constraint convexity requires log-concavity (Prékopa)

The feasible region $\{\mathbf{x} : \Pr[\mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x} \le b] \ge 1-\beta\}$ is convex if $\boldsymbol{\xi}$ has a *log-concave* density (Gaussian, uniform on a convex set, exponential), the basis for Theorem 12.3's SOC result, but can be *non-convex* for non-log-concave (e.g. multimodal) $\boldsymbol{\xi}$, in which case SOC reformulations may *exclude genuinely feasible points*, not merely approximate conservatively.

Proof sketch. Prékopa's (1971) theorem on log-concave measures shows the relevant super-level set is convex for log-concave $\boldsymbol{\xi}$ via marginalization properties specific to that class; no such guarantee holds for non-log-concave (e.g. multimodal) densities. ■

Example — A bimodal counterexample

Let $\boldsymbol{\xi} \sim \frac{1}{2}\mathcal{N}(-10, 1) + \frac{1}{2}\mathcal{N}(10, 1)$ (bimodal, non-log-concave). For $\Pr[\xi x \le 5] \ge 0.95$ with $x > 0$: $\Pr[\xi \le 5/x]$ jumps from ≈ 0 (for $5/x < -10$) to ≈ 0.5 (for $-10 < 5/x < 10$)

to ≈ 1 (for $5/x > 10$), so the constraint requires $5/x > 10 \iff x < 0.5$. The feasible set's shape across x can exhibit disconnections or curvature that a single-SOC-region reformulation cannot represent – a genuine modelling failure, not mere conservatism, when applied to multimodal $\boldsymbol{\xi}$.

Historical notes. Chance-constrained programming was introduced by Charnes and Cooper (1959, "Chance-Constrained Programming," *Management Science*), another product of the remarkably fertile 1950s operations-research period (alongside Dantzig's 1955 two-stage SP, Wald's 1945 maximin, and Scarf's 1958 moment-based newsvendor – all discussed in this volume's historical notes, and all within a 14-year span). The Gaussian-SOC reformulation (Theorem 12.3) follows directly from Charnes-Cooper's original analysis for the Gaussian case; the SOC *terminology* and its connection to the broader conic-programming framework of Part 9 is, naturally, much more recent (1990s, alongside the SOCP solvers that made the reformulation computationally practical at scale). The DRO-chance-constraint unification (this section's closing remark) is correspondingly recent, dating to the 2010s convergence of the DRO and CCP literatures discussed throughout Part 10's historical notes.

Theorem: the CVaR approximation to a chance constraint is conservative and asymptotically tight

For a loss $Z(\mathbf{x}) = \mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x} - b$ and chance constraint $\Pr[Z(\mathbf{x}) \le 0] \ge 1 - \beta$, the CVaR-based sufficient condition $\mathrm{CVaR}_\beta(Z(\mathbf{x})) \le 0$ implies the chance constraint (conservative direction, by $\mathrm{VaR}_\beta(Z) \le \mathrm{CVaR}_\beta(Z) \le 0 \implies \Pr[Z > 0] \le \beta$), and the two constraints coincide exactly (zero conservatism gap) in the special case that the conditional distribution of $Z(\mathbf{x})$ given $Z(\mathbf{x}) \ge \mathrm{VaR}_\beta(Z(\mathbf{x}))$ is degenerate (a point mass at the VaR threshold itself) – i.e. the tail beyond the β-quantile has no spread.

Proof. The implication direction is immediate from the definitions ($\mathrm{CVaR}_\beta \ge \mathrm{VaR}_\beta$ always, by definition of CVaR as a tail-conditional expectation dominating the corresponding quantile). The tightness claim follows because $\mathrm{CVaR}_\beta(Z) = \mathrm{VaR}_\beta(Z)$ exactly when the tail distribution beyond the VaR threshold is degenerate (no additional tail spread to average over), in which case the CVaR constraint ≤ 0 and the VaR/chance constraint ≤ 0 become literally the same inequality. ■

Example — Quantifying the conservatism gap for a Gaussian loss

For $Z(\mathbf{x}) \sim \mathcal{N}(\mu, \sigma^2)$ (as in Section 12.3's Gaussian chance-constraint reformulation), $\mathrm{VaR}_\beta = \mu + \sigma\Phi^{-1}(1-\beta)$ and $\mathrm{CVaR}_\beta = \mu + \sigma\,\phi(\Phi^{-1}(1-\beta))/\beta$ (Section 12.3's own worked CVaR formula). For $\beta = 0.05$: VaR factor $= 1.645$, CVaR factor ≈ 2.06 (Section 12.3's own worked example) – a conservatism gap of approximately 0.42 standard deviations, meaning the CVaR-based reformulation requires $\bar{a}^\top \mathbf{x} + 2.06\sigma_{\mathbf{x}} \le b$ where the exact chance constraint only requires $\bar{a}^\top \mathbf{x} + 1.645\sigma_{\mathbf{x}} \le b$, a strictly more demanding (more conservative) requirement on $\mathbf{x}$ by this margin. The Gaussian case is, in fact, close to the *worst* case for CVaR-vs-VaR conservatism among common light-tailed distributions, since the Gaussian's smoothly decaying tail gives substantial extra weight to the conditional tail expectation relative to the quantile alone.

When CVaR conservatism is small. For loss distributions with a sharply truncated or near-degenerate tail beyond the β-quantile (e.g. a bounded-support loss whose density vanishes rapidly just past VaR_β), the CVaR-VaR gap shrinks toward zero, recovering near-exact chance-constraint enforcement via the (always convex, always LP-representable per Section 12.1)

CVaR surrogate – giving a practical diagnostic: examine the *shape* of the loss distribution's upper tail (not just its overall variance) before deciding whether the CVaR approximation's conservatism is likely to be a major or minor practical cost for a given application, complementing the purely distributional (log-concavity) criterion of Theorem 12.3 in the main text.

Historical notes (addendum). The systematic use of CVaR as a tractable conservative surrogate for chance constraints, building on Rockafellar and Uryasev's (2000) original CVaR-optimization paper (already cited in this volume's bibliography), was developed specifically for the chance-constraint setting by Nemirovski and Shapiro (2006, "Convex Approximations of Chance Constrained Programs"), whose paper provides the general (non-Gaussian) conservatism bounds that the Gaussian-specific calculation above instantiates – their broader contribution, the systematic comparison of multiple convex chance-constraint surrogates (including but not limited to CVaR) by their relative conservatism, remains the standard reference for choosing among tractable chance-constraint approximations in non-log-concave settings.

12.4 Multi-Stage Stochastic Programming and Decision Trees

Overview

Multi-stage stochastic programming extends the two-stage model of Section 12.1 to $T > 2$ decision epochs, with the uncertainty realised progressively and represented by a *scenario tree*: decisions at stage t may depend on the (partial) history of $\boldsymbol{\xi}$ observed up to t, but not on the future – the formal *non-anticipativity* constraint.

Definition: Scenario tree and non-anticipativity

A scenario tree is a tree with root node n_0 (stage 0), and each node n at stage t has children at stage $t+1$ representing possible realisations of $\boldsymbol{\xi}_{t+1}$ given the history up to n. Each leaf (stage-T node) corresponds to a full scenario path with probability $p_{\text{leaf}} = \prod(\text{branch probabilities along the path})$. *Non-anticipativity*: the decision $\mathbf{x}_t(n)$ at node n depends only on the history up to n (the path from root to n), not on which child/descendant will be realised.

The multi-stage LP. Writing $\mathbf{x}_n$ for the decision vector at node n and $\text{anc}(n)$ for the parent of n,

$$\min_{\{\mathbf{x}_n\}} \sum_{n} p_n \, \mathbf{c}_n^{\top} \mathbf{x}_n \quad \text{s.t.} \quad \mathbf{T}_n \mathbf{x}_{\text{anc}(n)} + \mathbf{W}_n \mathbf{x}_n = \mathbf{h}_n \;\; \forall n, \quad \mathbf{x}_n \geq 0,$$

where p_n is the probability of reaching node n (sum of leaf probabilities below n). This is a single large-scale LP (the "deterministic equivalent" or extensive form) whose size grows *exponentially* in T for a tree with bounded branching factor b: $O(b^T)$ leaves.

Theorem (curse of dimensionality and decomposition). The number of scenarios in a tree with T stages and branching factor b at each stage is b^{T-1} (counting from stage 1), so the extensive-form LP has $O(b^{T-1})$ blocks of variables/constraints. However, the tree's structure – each node's constraints link only to its parent – means the LP has a *nested block-angular* structure, amenable to decomposition methods (nested Benders / stochastic dual dynamic programming, SDDP) whose per-iteration cost scales *linearly* (not exponentially) in the number of nodes, by solving a sequence of small subproblems coupled via "cut" (Benders cut) exchanges between adjacent stages.

Example — Tree size for a realistic planning horizon

$T = 12$ stages (e.g. monthly planning over a year), $b = 3$ branches per stage (e.g. low/medium/high demand). Number of scenarios $= 3^{11} \approx 1.77 \times 10^5$. Each scenario's LP block might have, say, 50 variables and 50 constraints, giving an extensive-form LP with $\approx 8.9 \times 10^6$ variables – at the edge of, but within reach of, modern LP solvers for $T = 12$. For $T = 24$ (two years monthly), $3^{23} \approx 9.4 \times 10^{10}$ scenarios – entirely intractable in extensive form, mandating decomposition (SDDP) or, as in Section 9.5, an affine-decision-rule-style approximation reducing the effective dimensionality from exponential-in-T to polynomial.

The SDDP algorithm sketch. SDDP builds piecewise-linear approximations $\underline{V}_t$ to value-to-go functions via forward passes (simulate, record trial decisions) and backward passes (solve each child scenario's subproblem, average dual prices into a new Benders cut). Convergence is monitored via the gap between a statistical upper bound and the deterministic lower bound $\underline{V}_1(\mathbf{x}_0)$.

Example — Cut accumulation in a 3-stage toy problem

Scalar state $x_t \in [0, 10]$, demand $\xi_t \in \{2, 5, 8\}$ each prob. $1/3$, independent across stages. After iteration 1, $\underline{V}_2$ gains one cut (averaging stage-3 dual prices over ξ_3). After iteration 2, $\underline{V}_2$ gains a second cut and $\underline{V}_1$ its first. With only 3 outcomes per stage, SDDP can converge in $O(3)$ iterations vs. the extensive form's $3^2 = 9$ leaf scenarios.

Stopping criteria and statistical bounds. Forward-pass costs give a CLT confidence interval (statistical upper bound); $\underline{V}_1(\mathbf{x}_0)$ is a guaranteed lower bound by convexity. SDDP stops when the interval's lower end approaches $\underline{V}_1(\mathbf{x}_0)(1 + \text{tol})$ – a primal-dual gap certificate analogous to those throughout convex optimization.

Connections. The exponential blow-up here is the multi-period analogue of the dimensionality issues encountered in Section 10.4's sample-complexity bounds: both reflect the fundamental difficulty of representing high-dimensional joint distributions (over time, or over $\mathbb{R}^d$) without structural simplification. The affine decision rules of Section 9.5, applied stage-by-stage ($\mathbf{x}_t = \mathbf{x}_t^0 + \sum_{\tau<t} \mathbf{Y}_{t,\tau}\boldsymbol{\xi}_\tau$), give one such polynomial-size approximation for multi-stage *robust* (rather than stochastic) problems; SDDP gives an exact (in the sample-average sense) but iterative alternative for the stochastic case, and the choice between the two – robust/affine vs. stochastic/SDDP – is precisely the robust-vs-stochastic comparison formalised in the final section, 12.5.

Theorem: value-function monotonicity across SDDP iterations

Let $\underline{V}_t^{(k)}$ denote the SDDP lower-bound approximation to V_t after k iterations. Then $\underline{V}_t^{(k)}(\mathbf{x}) \le \underline{V}_t^{(k+1)}(\mathbf{x}) \le V_t(\mathbf{x})$ for all $\mathbf{x}$ and all k – the lower bounds are monotonically non-decreasing in k and always bounded above by the true value function, so $\underline{V}_1^{(k)}(\mathbf{x}_0) \nearrow$ as $k \to \infty$, with the limit equal to $V_1(\mathbf{x}_0)$ under standard regularity conditions (SDDP convergence).

Proof sketch. $\underline{V}_t^{(k+1)} = \max(\underline{V}_t^{(k)}, \text{new cut at iteration } k+1)$ (each new Benders cut is added to the existing lower envelope, which can only increase the pointwise maximum, hence $\underline{V}_t^{(k+1)} \ge \underline{V}_t^{(k)}$). Each individual cut is a valid lower bound on V_t (Benders cuts are derived from convex duality and are, by construction, supporting hyperplanes of the true convex V_t, hence $\le V_t$ everywhere), so $\underline{V}_t^{(k)} = \max(\text{cuts so far}) \le V_t$ for all k. Monotone-bounded

sequences converge; convergence to V_t itself (rather than some smaller limit) requires that the forward-pass sampling visits a dense-enough set of $\mathbf{x}$ values over many iterations (an ergodicity-type condition), the technical heart of SDDP convergence proofs. ■

Example — Tracking convergence numerically

For the 3-stage toy problem (Section 12.4's case study), suppose $\underline{V}_1^{(1)}(\mathbf{x}_0) = 40$, $\underline{V}_1^{(2)}(\mathbf{x}_0) = 44$, $\underline{V}_1^{(3)}(\mathbf{x}_0) = 46.5$, $\underline{V}_1^{(4)}(\mathbf{x}_0) = 47.2$, with statistical upper bounds (from forward-pass sample costs) at iterations 1-4 of, say, $52, 49, 48, 47.6$ (also generally decreasing as the policy improves, though *not* guaranteed monotone the way the lower bound is, since it's a Monte Carlo estimate with sampling noise). The gap (upper − lower) shrinks: $12, 5, 1.5, 0.4$ – a typical SDDP convergence pattern (rapid initial gap closure, slowing as the value function approximation refines), with the lower bound's strict monotonicity (Theorem above) providing the *algorithmic* guarantee that progress, once made, is never lost – in contrast to, e.g., gradient-based methods where objective values can temporarily worsen before improving.

Remarks and extensions. The scenario-tree formulation of multi-stage stochastic programming implicitly assumes a *discrete-time*, finite-horizon structure. Continuous-time analogues (stochastic control, covered in the companion volumes on stochastic processes) replace the scenario tree by a filtration of a continuous-time stochastic process (e.g. a diffusion), and the multi-stage LP of this section becomes a *Hamilton-Jacobi-Bellman* (HJB) partial differential equation for the value function $V(t, \mathbf{x})$. The SDDP algorithm's value-function approximations $\underline{V}_t(\mathbf{x})$ can be seen as a (piecewise-linear, discrete-time, discrete-state) discretization of the continuous HJB value function – the same value-function concept appears in both discrete and continuous formulations, with SDDP and finite-difference/finite-element HJB solvers as the respective discrete and continuous numerical methods.

A practical remark on rolling-horizon implementation: in practice, multi-stage stochastic programs are rarely solved once and then executed open-loop for all T stages. Instead, a *rolling-horizon* (or *model-predictive-control*) approach re-solves the multi-stage problem at each stage t, using the *realised* $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{t-1}$ (rather than the scenario-tree's hypothetical branches) as the conditioning history, and a freshly-generated scenario tree for the remaining stages $t, \dots, T$. This re-solving strategy is computationally more expensive (solving T multi-stage problems instead of 1) but is far more robust to the scenario tree's inevitable misspecification of the true (continuous, high-dimensional) uncertainty – a practical reconciliation between the elegant theory of this section and the messiness of real deployment, paralleling the practical-vs-theoretical tension noted in Section 11.5's remarks on info-gap theory.

Theorem: SDDP cut-sharing across structurally identical stages

If stages t, t' have identical cost structure, dynamics, and uncertainty distribution, Benders cuts computed for $\underline{V}_t$ are also valid lower bounds for $\underline{V}_{t'}$ – cuts can be *shared* across structurally identical stages, reducing the number of distinct value-function approximations from T to the number of *distinct stage-types*.

Proof. A cut $\underline{V}_t(\mathbf{x}) \geq \alpha + \boldsymbol{\beta}^\top \mathbf{x}$ is derived from stage-$(t+1)$ dual prices; if stage $t'+1$ has identical structure, the same derivation yields the same valid lower bound for $V_{t'}$. ■

Example — Periodic reservoir management

For a multi-year extension of the hydroelectric case study (e.g. $T = 48$ months $= 4$ years $\times 12$, but only 4 *seasonal* stage-types), cuts from any winter stage's backward pass contribute to *all* winter stages – reducing distinct value-function approximations from 48 to 4, a $12\times$ reduction. This cut-sharing is a major practical efficiency technique in real SDDP implementations for periodic/seasonal problems, essential for the multi-year horizons (capturing drought cycles etc.) often required in practice.

Historical notes. Multi-stage stochastic programming's scenario-tree formulation generalizes Dantzig's 1955 two-stage framework (Section 12.1's notes) in a relatively straightforward conceptual step, but its *computational* tractability for realistic T required the nested-decomposition / SDDP algorithmic development of Pereira and Pinto (1991, "Multi-Stage Stochastic Optimization Applied to Energy Planning"), motivated specifically by the Brazilian hydroelectric system planning problem described in this section's case study. SDDP's subsequent three-decade track record of large-scale industrial deployment (Brazil, and later adopted by grid operators in Norway, Colombia, and elsewhere) makes it one of the most practically successful algorithms to emerge from the stochastic programming literature – a notable contrast with, e.g., the still-largely-academic status (as of this writing) of some of the more recent DRO and info-gap frameworks discussed elsewhere in this volume, illustrating that practical adoption timescales for theoretically elegant frameworks can vary enormously (SDDP: roughly a decade from theory to large-scale deployment; some DRO variants: still pre-widespread-deployment three decades after Scarf's 1958 precursor).

Theorem: nested CVaR objectives preserve SDDP's convexity and cut-validity structure

Replacing the risk-neutral multi-stage objective $\mathbb{E}[\sum_t c_t^\top \mathbf{x}_t]$ by a *nested* (time-consistent) risk measure $\rho_1(c_1^\top \mathbf{x}_1 + \rho_2(c_2^\top \mathbf{x}_2 + \cdots))$, where each $\rho_t(\cdot) = (1-\lambda)\mathbb{E}_t[\cdot] + \lambda\,\mathrm{CVaR}_{\beta,t}(\cdot)$ is a fixed convex combination of expectation and conditional CVaR at stage t, preserves both the convexity of each stage's value function $V_t(\mathbf{x})$ and the validity of Benders-type cuts generated from each stage's (now risk-adjusted) dual prices – SDDP, with this modification to the backward-pass cut-generation step, remains a convergent algorithm for the risk-averse nested-CVaR objective.

Proof sketch. $\mathrm{CVaR}_\beta(\cdot)$ is itself a coherent (in particular, convex) risk measure with an LP representation (Section 12.1's remark, generalizing Rockafellar–Uryasev); a convex combination of the (linear, hence convex) expectation operator and a convex risk measure remains convex, and composing convex stage-wise risk measures in the nested (nested, not just expectation-of-CVaR) fashion preserves overall convexity by an induction argument backward through the stages (a standard result in the dynamic/nested risk-measure literature, e.g. Ruszczyński 2010). Since each ρ_t retains an LP/SOCP representation, the dual prices needed for Benders cuts remain well-defined and the cut-validity argument of Section 12.4's main-text SDDP cut-monotonicity theorem carries through essentially unchanged, with the nested risk measure's dual variables replacing the ordinary expectation's. ■

Example — Risk-averse vs. risk-neutral SDDP on the reservoir case study

Returning to Section 12.4's periodic hydroelectric reservoir example: a risk-neutral SDDP run minimizes expected cost across all hydrological scenarios, potentially accepting a small probability of a very costly drought-driven supply shortfall in exchange for lower expected cost. A nested-CVaR run with $\lambda = 0.5$, $\beta = 0.1$ at each

stage explicitly penalizes the worst 10% of cost outcomes at every stage (not just the terminal stage), typically yielding a policy with a higher expected cost but a substantially lower 90th-percentile worst-case cost – the same risk-neutral-vs-robust trade-off encountered throughout this volume (e.g. Section 12.5's newsvendor RO-vs-SP comparison), now realized *within* the multi-stage SDDP framework itself rather than as a choice between entirely separate RO and SP paradigms.

Connection to Section 12.5's unifying table. This risk-averse SDDP variant occupies an instructive position relative to Section 12.5's 2×2 unifying table: it is not literally any of the four named corners (RO, DRO, CCP, SP) but rather a *risk-measure-parametrized* family interpolating between SP ($\lambda = 0$, pure expectation) and an increasingly RO-like worst-case orientation as $\lambda \to 1$ and $\beta \to 0$ simultaneously (since $\text{CVaR}_{\beta\to 0} \to$ ess sup, recovering a worst-case-style objective, per the CVaR-limit identities used throughout Section 12.5). Nested-CVaR SDDP can thus be seen as a third continuous parameter (alongside ϵ and β in Section 12.5's table) for navigating the same underlying landscape, this time specifically in the multi-stage setting where the scenario-tree's full path structure (not just its marginal distribution at each stage) is relevant to risk attitude.

Historical notes (addendum). Time-consistent (nested) risk measures for multi-stage stochastic programming were developed by Ruszczyński and Shapiro (2006, "Conditional Risk Mappings") specifically to address a subtlety glossed over by naively applying a single risk measure to the *terminal* cost only (which, as several authors showed, can induce time-*inconsistent* optimal policies – a policy optimal when computed at stage 1 may no longer appear optimal when re-evaluated at stage 2 given realized information); the nested construction used in this section's theorem is specifically designed to avoid this pathology, at the cost of the more elaborate stage-by-stage risk-measure composition described above.

12.5 Robust vs Stochastic: A Formal Comparison

Overview

This closing section directly compares the robust optimization (Part 9, worst-case over a set $\mathcal{U}$) and stochastic programming (Sections 12.1, 12.4, expectation over a distribution P) paradigms, formalising when each is preferable and how the distributionally robust (Part 10) and chance-constrained (Section 12.3) frameworks interpolate between them.

Theorem: RO as a limiting case of CCP

The chance constraint $\Pr[\mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x} \le b] \ge 1-\beta$ converges, as $\beta \to 0$, to the constraint $\mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x} \le b$ *almost surely* (with respect to the support of $\boldsymbol{\xi}$'s distribution). If $\boldsymbol{\xi}$ has bounded support $\mathcal{U}$, "almost surely on $\mathcal{U}$" is precisely the robust constraint $\sigma_{\mathcal{U}}(\mathbf{x}) \le b$ of Section 9.1. Thus $\beta = 0 \iff$ RO, and $\beta = 1 \iff$ no constraint at all (only the nominal/expected-value version matters) – RO and "no robustness" are the two endpoints of the CCP reliability spectrum $\beta \in [0,1]$.

A unifying table. The four frameworks of Parts 9–12 can be organised by (i) what is uncertain (a parameter $\mathbf{a}$, or a distribution P) and (ii) what guarantee is sought (almost-sure, or in-expectation/probabilistic):

	Almost-sure guarantee	**Probabilistic/expectation guarantee**
Uncertain parameter a $\in \mathcal{U}$	Robust Optimization (Part 9)	Chance-Constrained Prog. (Sec. 12.3)
Uncertain distribution $P \in \mathcal{P}$	DRO, worst-case (Part 10)	Stochastic Prog. (Sec. 12.1, 12.4)

DRO (bottom-left) is in fact the "robust" version of stochastic programming (bottom-right), exactly as RO (top-left) is the "robust" version of CCP (top-right) – the same row-vs-column logic of Theorem 12.5's $\beta \to 0$ limit applies columnwise: as the ambiguity set $\mathcal{P} \to \{\hat{P}_N\}$ (a singleton), DRO collapses to ordinary stochastic programming with $P = \hat{P}_N$.

Example — Numerical comparison on the newsvendor problem

Recall Section 12.1's newsvendor with $c = 2, h = 1, p = 5$, demand $\xi \sim \text{Uniform}(0, 100)$ (true distribution). **Stochastic-optimal** (Sec. 12.1): $x^*_{\text{SP}} = 50$ (median), expected cost $\approx \mathbb{E}[cx + Q(x, \xi)]$; computing $\mathbb{E}[Q(50, \xi)] = \frac{1}{100}\int_0^{50} 1 \cdot (50 - \xi)d\xi + \frac{1}{100}\int_{50}^{100} 5 \cdot (\xi - 50)d\xi = \frac{1}{100}(1250) + \frac{1}{100}(6250) = 12.5 + 62.5 = 75$; total $= 2(50) + 75 = 175$. **Robust** ($\mathcal{U} = [0, 100]$, worst case $\xi = 100$ if $x < 100$, contributing $5(100 - x)$, or $\xi = 0$ if $x > 0$, contributing $1 \cdot x$; worst case is $\max(5(100 - x), x)$, minimized at $5(100 - x) = x \Rightarrow x = 500/6 \approx 83.3$, worst-case recourse cost ≈ 83.3, total $\approx 2(83.3) + 83.3 = 250$). The robust solution ($x \approx 83.3$, total worst-case cost 250) over-orders relative to the stochastic-optimal ($x = 50$, expected cost 175) – protecting against the worst case ($\xi = 100$, a probability-zero event under continuous uniform but the supremum of the support) costs an extra 75 in expected terms, the "price of robustness" (Section 9.4) made concrete in a stochastic setting. A DRO solution with a moderate ambiguity radius would interpolate between $x = 50$ and $x \approx 83.3$, with the radius controlling exactly where on this spectrum the solution lies.

The role of $\mathcal{U}$ vs. $\mathcal{P}$ in practice. If historical data constrains $\boldsymbol{\xi}$'s *support* but not its distribution, RO/DRO (left column) suffice with only support knowledge. If a distribution can be confidently fitted but small-β tail guarantees matter, CCP/SP (right column) with the fitted distribution are natural, with exact-vs-DRO-CCP governed by confidence in distributional *shape*.

Example — A combined RO+SP example

First-stage $\mathbf{c} \in \mathcal{U}_{\text{box}}$, second-stage demand $\boldsymbol{\xi} \sim P$ (well-estimated): $\min_{\mathbf{x}} \max_{\mathbf{c} \in \mathcal{U}_{\text{box}}} [\mathbf{c}^\top \mathbf{x} + \mathbb{E}_{\boldsymbol{\xi} \sim P}[Q(\mathbf{x}, \boldsymbol{\xi})]]$. By Eq. (9.1.box), $\max_{\mathbf{c}} \mathbf{c}^\top \mathbf{x} = \bar{\mathbf{c}}^\top \mathbf{x} + \boldsymbol{\rho}^\top |\mathbf{x}|$, a deterministic term. The combined problem is $\min_{\mathbf{x}} [\bar{\mathbf{c}}^\top \mathbf{x} + \boldsymbol{\rho}^\top |\mathbf{x}| + \mathbb{E}_{\boldsymbol{\xi} \sim P}[Q(\mathbf{x}, \boldsymbol{\xi})]]$ – the price-of-robustness penalty simply *adds* to the recourse term; additively separable uncertainty types combine by addition.

Looking ahead. The robust-vs-stochastic dichotomy, resolved via the unifying table, exemplifies this volume's lesson: distinct frameworks are related by precise limits ($\epsilon \to 0$, $\beta \to 0$, $\Gamma \to$ singleton) to a common core – support functions, Hölder/conic duality, convex value functions. Volume IV asks how this toolkit behaves when decisions and uncertainty are *network states*, where local robustness interacts with global propagation dynamics producing qualitatively new phenomena (cascades, tipping points).

Synthesis of Volume III. The unifying table above is the capstone of this volume: every framework developed across Parts 9–12 – uncertainty-set geometry (Part 9), ambiguity sets over distributions (Part 10), non-additive decision criteria (Part 11), and explicit probabilistic/multi-stage models (Part 12) – occupies a definite position in a two-dimensional landscape indexed

by *what* is uncertain and *what kind of guarantee* is required, with explicit limiting relationships ($\beta \to 0$, $\mathcal{P} \to \{\hat{P}\}$, $\Gamma \to \{\text{all priors}\}$, $\epsilon \to 0$) connecting adjacent cells. Volume IV (Cascading Failures & Network Resilience) builds on this foundation by embedding these single-decision frameworks into the dynamics of interconnected systems, where a "decision" at one node propagates – robustly, stochastically, or adversarially – through a network.

Theorem: the unifying table's diagonal is self-dual

In the 2×2 table of Section 12.5, the "diagonal" entries – (uncertain parameter, almost-sure) = RO, and (uncertain distribution, expectation) = SP – are related by a formal *duality*: the RO problem $\min_{\mathbf{x}} \max_{\mathbf{a}\in\mathcal{U}} \ell(\mathbf{x},\mathbf{a})$ and the SP problem $\min_{\mathbf{x}} \mathbb{E}_{\boldsymbol{\xi}\sim P}[\ell(\mathbf{x},\boldsymbol{\xi})]$ are the $\beta = 0$ and "β = no constraint" instances, respectively, of the *same* family $\min_{\mathbf{x}} \mathrm{CVaR}_{\beta}^{\mathcal{U} \text{ or } P}[\ell(\mathbf{x},\cdot)]$ as $\beta \to 0$ (CVaR_0 = essential supremum = $\max_{\mathbf{a}\in\mathcal{U}}$ for $\mathcal{U}$ = support) and $\beta \to 1$ ($\mathrm{CVaR}_1 = \mathbb{E}$), with the off-diagonal entries (DRO, CCP) interpolating at intermediate $\beta \in (0,1)$ *and* intermediate ambiguity-set size.

Proof sketch. $\mathrm{CVaR}_\beta(Z) \to \operatorname{ess\,sup}(Z)$ as $\beta \to 0$ and $\mathrm{CVaR}_\beta(Z) \to \mathbb{E}[Z]$ as $\beta \to 1$ are standard limiting properties of CVaR (following from its definition as a tail-conditional expectation: as the tail probability $\beta \to 0$, the "tail" shrinks to just the supremum; as $\beta \to 1$, the "tail" expands to the whole distribution). Applying these limits to $Z = \ell(\mathbf{x},\boldsymbol{\xi})$ with $\boldsymbol{\xi}$ ranging over $\mathcal{U}$ (RO case, with P = any distribution supported on $\mathcal{U}$, so ess sup = $\max_{\mathcal{U}}$) or $\boldsymbol{\xi} \sim P$ (SP case) gives the stated correspondences. ■

Example — A single CVaR-based implementation spanning the whole table

A risk management system implements a single solver for $\min_{\mathbf{x}} \mathrm{CVaR}_{\beta}^{\mathcal{P}}[\ell(\mathbf{x},\boldsymbol{\xi})]$ where $\mathcal{P}$ is a Wasserstein ball of radius ϵ around $\hat{P}_N$ (a DRO-CVaR hybrid, itself representable via the duality results of Sections 10.2-10.3 combined with the CVaR-LP representation of Section 12.1). Setting $(\epsilon,\beta) = (0,1)$ recovers plain SP (bottom-right cell); $(\epsilon,\beta) \to (\infty,0)$ recovers RO over the full support (top-left cell, in the limit where $\mathcal{U}$ = everything, an admittedly degenerate/uninteresting RO instance, but illustrating the limit's existence); finite $(\epsilon,\beta=0)$ recovers DRO (bottom-left); finite $(\epsilon=0,\beta\in(0,1))$ recovers ordinary CCP/CVaR-constrained SP (top-right, via the CVaR-VaR relationship of Theorem 12.3's box). A *single* two-parameter family $(\epsilon,\beta) \in [0,\infty)\times[0,1]$ thus spans the entire unifying table (and its continuous interior) – the practical, implementation-level realization of the "continuous two-parameter family" remark closing Section 12.5's main text, and a fitting final synthesis for Volume III: every framework in this volume is a corner, edge, or interior point of a single two-dimensional family, itself built from the support-function and CVaR machinery introduced across Parts 9 and 12.

Remarks and extensions. The unifying 2×2 table of this section, while a useful organizing device, necessarily simplifies a more continuous reality: the four "cells" are connected not just at their shared boundaries ($\beta = 0$, $\mathcal{P} = \{\hat{P}\}$, etc.) but via a richer web of intermediate frameworks not explicitly named in the table – e.g. DRO chance constraints (Section 12.3's closing remark) occupy a position "between" the bottom-left (DRO) and top-right (CCP) cells, simultaneously robust to distributional ambiguity *and* probabilistic (rather than almost-sure) in their guarantee. A more accurate picture replaces the 2×2 table with a continuous two-parameter family, indexed by (i) the "size" of the ambiguity set (ϵ or $|\Gamma|$, with 0 = singleton and ∞ = all distributions) and (ii) the reliability level $\beta \in [0,1]$, with the four named frameworks occupying the four corners $(\epsilon,\beta) \in \{0,\infty\}\times\{0,1\}$ of this continuous square.

As a closing remark for the volume as a whole: the mathematical machinery developed here – support functions and dual norms (Part 9), conic and transport duality (Part 10), capacities and sets-of-priors (Part 11), and value functions and Benders decomposition (Part 12) – recurs, often under different names and in different notation, throughout the broader landscape of applied mathematics: in robust control theory (H_∞ control is, formally, a robust-optimization problem over a frequency-domain uncertainty set), in robust statistics (Section 11.4's remarks), in online learning and regret minimization (Section 11.2's minimax regret, generalised to sequential/adaptive settings), and in the optimization-landscape and generalization-theory chapters of the companion Mathematics of Neural Networks series (Section 10.5's gradient-regularization connection). Readers who internalise the support-function/duality core of this volume will find it a remarkably portable piece of mathematical equipment, recognisable – once one knows to look for it – across an unusually wide swath of quantitative disciplines.

Theorem: each cell of the unifying table has both a tractable and a hard regime

RO (top-left) is LP/SOCP for polyhedral/ellipsoidal $\mathcal{U}$ (Theorem 9.2); DRO (bottom-left) is LP/SOCP/SDP depending on ambiguity type (Part 10) but can require unbounded Lasserre degree for high-degree ℓ (Section 10.3); CCP (top-right) is SOCP for log-concave distributions (Theorem 12.3) but can be non-convex otherwise (this section's earlier theorem); SP (bottom-right) is LP for finite scenarios but the extensive form is exponential in T without decomposition (Section 12.4). **No cell is uniformly easiest** – each has a tractable convexity-dependent sub-case and a hard general case.

Proof sketch. Synthesis of Theorem 9.2 (RO), Section 10.3's Lasserre remarks (DRO), this section's Prékopa-based theorem (CCP), and Section 12.4's exponential-extensive-form discussion (SP). ■

Example — A practitioner's decision checklist, synthesised

(1) Is ℓ convex in $\mathbf{x}$? If not, none of Parts 9-12's results apply directly. (2) Is the uncertainty's *support* ($\to$RO/CCP, top row) or *distribution* ($\to$DRO/SP, bottom row) known? (3) Almost-sure ($\to$left column) or probabilistic ($\to$right column) guarantee – and if probabilistic, is the distribution log-concave (Theorem above)? If not, consider SAA or DRO-CCP instead of SOC. (4) For multi-stage, check b^{T-1}: if large, plan for SDDP (Section 12.4) or ADR (Section 9.5) from the outset. This checklist operationalises the unifying table's classification logic into concrete, checkable questions – the practical culmination of Volume III, the point of transition to volume-specific solution techniques (LP/SOCP/SDP solvers, SDDP) referenced throughout this volume's historical notes.

Historical notes. The explicit "unifying table" presented in this closing section is, to the authors' knowledge, not a standard textbook fixture (most treatments present RO, DRO, CCP, and SP as separate chapters or even separate books, without an explicit organizing meta-structure of this kind) – it represents this volume's attempted synthesis rather than a result with its own independent citation history. However, each of the *individual* connections in the table (RO as $\beta \to 0$ CCP; DRO as worst-case-P SP; the Γ-minimax/DRO identification of Section 11.4's notes) has appeared, in various forms, across the literatures surveyed in this volume's historical notes – the synthesis offered here is one of *assembly* rather than *discovery*. Readers interested in primary sources for the individual connections are referred to the historical notes of the relevant earlier sections; readers interested in the broader question of

why these connections exist (the support-function/duality core identified in Section 12.5's main text and in this volume's closing remarks) are referred to Rockafellar's *Convex Analysis* (1970) and its successors as the deepest common root from which essentially all of Part 9-12's mathematical machinery, traced back far enough, descends.

PART 13

Synthesis, Open Problems, and Connections to Adjacent Fields

A capstone treatment situating the volume's four frameworks within broader mathematical landscapes: robust statistics, adversarial machine learning, financial risk management, and network resilience.

Theorem: the meta-level choice between RO and SP is itself a decision under (Knightian) uncertainty

Given a sample of N historical observations of the uncertain parameter, with no further information about whether the data-generating process is well approximated by a fixed, fully-known distribution (favouring the SP route) or is better treated as merely support-constrained (favouring the RO route), the choice of *which framework to deploy* is formally a decision problem under Knightian uncertainty over "which modelling paradigm is correct" – and can itself be approached via the very decision criteria (maximin, Hurwicz, Γ-minimax) developed in Part 11, applied at the meta-level of model selection rather than at the level of the original decision variable $\mathbf{x}$.

Proof sketch. This is an immediate reframing rather than a technical result: let $\mathbf{x}^\star_{\text{SP}}$ and $\mathbf{x}^\star_{\text{RO}}$ be the two frameworks' respective solutions, and consider the two possible "states of the world" (the data-generating process is well-described by the fitted distribution; or it is better described as merely support-constrained, with the fitted distribution being a poor model of tail behaviour). A decision-maker uncertain between these two states faces exactly the structure of a two-state, two-action decision problem under Knightian uncertainty over which the maximin/Hurwicz/Γ-minimax machinery of Part 11 directly applies, now with "which $\mathbf{x}^\star$ to deploy" as the meta-decision and "which paradigm is correct" as the unknown state. ■

EXAMPLE — A meta-level minimax illustration

Suppose the SP-optimal $\mathbf{x}^\star_{\text{SP}} = 50$ performs excellently (cost 175) if the fitted Uniform$(0, 100)$ distribution is correct, but poorly (cost 350) if the true tail is much heavier than fitted and a rare extreme demand realization occurs, under the alternative "misspecified-tail" state; conversely $\mathbf{x}^\star_{\text{RO}} \approx 83.3$ performs moderately under both states (cost 260 if SP's distributional assumption holds, cost 270 under the misspecified-tail state, since it was already designed to hedge against the support's extreme). Since these are *costs* (lower is better), the relevant worst-case criterion is *minimax*, not maximin: a meta-level minimax comparison evaluates the worst-case cost of each framework, $\max(175, 350) = 350$ for SP versus $\max(260, 270) = 270$ for RO, and selects whichever framework minimizes this worst case – here, RO, since $270 < 350$. A meta-level *minimax-regret* comparison instead weighs each framework's worst-case *regret* relative to the best achievable cost in each state (175 in the good state, 260 in the bad state): SP's worst regret is $\max(175 - 175,\ 350 - 260) = 90$, while RO's worst regret is $\max(260 - 175,\ 270 - 260) = 85$ – so by the same numbers, minimax-regret *also* favours RO here, but by a much narrower margin (85 vs. 90) than the minimax-cost comparison's margin (270 vs. 350), illustrating that even when two meta-level criteria happen to agree on which framework to deploy, they can disagree sharply on *how decisively* – a distinction relevant whenever a decision-maker must also communicate their confidence in the chosen framework, not merely the choice itself.

A practical resolution: nested robustness. In practice, this meta-level uncertainty is often resolved not by a single global choice of framework but by the DRO and CCP frameworks

of Parts 10 and 12 themselves, which are explicitly designed as intermediate points between pure RO and pure SP (Section 12.5's unifying table) – in this sense, DRO and DRO-CCP hybrids can be read as a *constructive*, rather than purely decision-theoretic, resolution of the meta-level RO-vs-SP uncertainty highlighted by this section's theorem: rather than choosing one paradigm via a meta-level maximin/regret calculation, the decision-maker selects an ambiguity-set *size* (ϵ or Γ) that directly encodes their meta-level confidence in the fitted distribution, smoothly interpolating between the two extremes rather than committing to either pole.

Historical notes (addendum). The recognition that model-selection itself constitutes a decision under uncertainty, amenable to the same formal apparatus used for the underlying decision problem, is a recurring theme in the foundations of statistics (visible already in Wald's 1950 framework, which treats the choice of decision rule and the choice of prior on comparable formal footing) but is rarely made explicit in the operations-research-oriented RO/SP literature, which more often presents the choice between frameworks as a matter of practitioner judgment or convention rather than as a formal decision problem in its own right – this section's explicit framing is intended to close that gap and to position the entire RO/DRO/CCP/SP landscape of this volume within the broader decision-theoretic apparatus of Part 11.

13.1 The Core Mathematical Thread: Support Functions and Duality

Overview

Part 13 does not introduce new material but re-examines the four frameworks of this volume – Robust Optimization (Part 9), Distributionally Robust Optimization (Part 10), Decision Under Deep Uncertainty (Part 11), and Stochastic Programming (Part 12) – from a unified mathematical vantage point, identifies the open research questions that each part's development left explicitly unresolved, and maps the connections to adjacent mathematical disciplines.

A recurring observation, first made in passing in Sections 9.3 and 10.3 and culminating in Section 12.5's unifying table, is that every framework in this volume derives its tractability from the same family of mathematical tools: *support functions* and *conjugate duality*. This section makes that connection explicit.

The support function $\sigma_{\mathcal{U}}(\mathbf{x}) = \sup_{\mathbf{a}\in\mathcal{U}} \mathbf{a}^\top \mathbf{x}$ is the central object of Part 9: it converts the nonlinear max-over-set operation into a function that is always convex in $\mathbf{x}$ (as a supremum of linears), whose properties are entirely determined by the geometry of $\mathcal{U}$. Box, ellipsoidal, and polyhedral uncertainty sets correspond, via their support functions, to the ℓ_∞-, ℓ_2-, and general-norm balls of the dual space – the geometry of $\mathcal{U}$ maps to the geometry of a norm ball via the conjugate-norm construction of Appendix A.

Part 10 (DRO) extends this from uncertainty over a *parameter* to uncertainty over a *distribution*: the Wasserstein ball's dual, derived in Section 10.3, is again a support function in the Kantorovich sense – the Lipschitz-function class defines the "geometry" of the Wasserstein ball exactly as the norm ball defines the geometry of ellipsoidal uncertainty sets in Part 9. The ϕ-divergence ambiguity sets of Section 10.1 use the conjugate of ϕ in the same role. Part 11 (Choquet integration, Γ-minimax) generalises still further: the lower expectation $\underline{E}_\Gamma[Z] = \inf_{P\in\Gamma} \mathbb{E}_P[Z]$ is the support function of the convex set Γ (of probability measures) evaluated in the direction Z. Even SDDP (Part 12) belongs: the value-function approximations $\underline{V}_t^{(k)}(\mathbf{x})$ are piecewise-linear support functions of their Benders-cut collections.

This unity is not accidental. It reflects the deep structure identified by Rockafellar (1970) and later Nesterov–Nemirovski (1994): *tractable convex optimization is, in large part, the theory of support functions and their calculus.* The frameworks of this volume are applications of that calculus to different objects (uncertain parameters, uncertain distributions, uncertain priors, uncertain multi-stage futures) – each framework being precisely the extension required when one additional layer of uncertainty is admitted.

Example — The support-function thread, traced through all four Parts

Part 9: $\max_{\mathbf{a}\in\mathcal{U}} \mathbf{a}^\top\mathbf{x} = \sigma_{\mathcal{U}}(\mathbf{x})$ (Definition 9.1.1). For $\mathcal{U} = \mathcal{U}_{\text{ell}}$: $\sigma_{\mathcal{U}}(\mathbf{x}) = \bar{\mathbf{a}}^\top\mathbf{x} + \left\|\mathbf{P}^\top\mathbf{x}\right\|_2$ (Theorem E.1). The robust constraint becomes a SOC constraint: support function of an ellipsoid is a Euclidean-norm expression. **Part 10**: $\sup_{Q:W_1(Q,\hat{P}_N)\le\epsilon} \mathbb{E}_Q[\ell] = \hat{\mathbb{E}}_N[\ell] + \epsilon\,\|\cdot\|_*$ (Theorem E.2). The Wasserstein ball's "support" in the direction ℓ is the empirical mean plus a dual-norm term. **Part 11**: $\underline{E}_\Gamma[Z] = \inf_{P\in\Gamma} \mathbb{E}_P[Z]$ (Section 11.4). For Γ a Wasserstein ball of priors: same expression as Part 10 with the role of "loss" and "prior" interchanged. **Part 12**: SDDP lower bounds $\underline{V}_t^{(k)}(\mathbf{x}) = \max(\text{Benders cuts so far})$ are pointwise maxima of linear functions, i.e. support functions.

13.2 Open Problems and Directions

Each part of this volume left specific technical problems unresolved, either because their resolution requires tools substantially beyond the volume's scope, or because they are genuinely open research questions. This section catalogues the most important.

From Part 9 (RO). Multi-stage adjustable robust optimization (Section 9.5) with affine decision rules is tractable; non-affine rules are generally NP-hard (Bertsimas and Goyal, 2012). The fundamental open question is: for which classes of uncertainty sets $\mathcal{U}$ and which classes of recourse functions $\mathcal{F}$ does the adjustable robust problem admit a polynomial-time exact or approximately-exact reformulation, beyond the known affine and ball-constrained special cases? Recent work on "non-anticipativity constraints" and "constraint-satisfying recourse" has made partial progress, but no general tractability characterization is known.

From Part 10 (DRO). The sample-complexity results of Section 10.4 give *in-sample* Wasserstein radius ϵ_N guarantees. Two related open questions are: (i) how should ϵ_N be set in high-dimensional problems where the curse of dimensionality makes W_1 estimation slow (the known $O(N^{-1/d})$ rate for d-dimensional distributions is exponentially slow in d), and (ii) for which loss classes do Wasserstein-DRO's out-of-sample guarantees remain sharp (tight) rather than loose? The interaction between the Wasserstein geometry (inherited from the metric d on Ξ) and the loss function's geometry is not yet fully understood in high dimension.

From Part 11 (Deep uncertainty). Info-gap theory's lack of a probabilistic foundation (explicitly noted in Section 11.5's remarks) makes it difficult to compare its robustness function $\hat{\alpha}(\mathbf{x}, r_c)$ against DRO or RO on an equal footing. A rigorous bridge – translating info-gap horizons of uncertainty into DRO ambiguity-set radii or RO uncertainty-set scales – remains an open modelling problem with significant practical implications for engineering applications (where info-gap is most used). Conversely, the axiomatic foundations of non-additive probability (Section 11.3's Choquet approach) are better understood than their computational counterparts: for which capacity classes does the Choquet integral $\int f\,d\nu$ admit a polynomial-time approximation when Ω is exponentially large (as in combinatorial optimization)?

From Part 12 (SP). SDDP's empirical convergence behaviour (Appendix C) is well-understood

practically but poorly bounded theoretically. The open question – whether SDDP has a polynomial worst-case iteration bound for some structurally natural class of multi-stage problems (periodic, Markovian, or with special stage-wise distributions) – has motivated a line of work on "regularized SDDP" and "level-function SDDP" that partially addresses it but does not resolve it in full generality. A related open problem is the integration of DRO ambiguity (Part 10) into multi-stage stochastic programs: DRO-SDDP (replacing the fixed stage-wise distribution by a Wasserstein ambiguity set) has been studied for two-stage problems but its extension to $T > 2$ stages raises subtle issues of time-consistency (whether the worst-case distribution at stage t can "change" at stage $t+1$) that are not yet fully resolved.

13.3 Connections to Adjacent Fields

Robust statistics. The min-max framework of Part 9 (minimizing worst-case loss over an uncertainty set) has a direct counterpart in robust statistics: Huber's (1964) minimax contamination model minimizes worst-case risk over a contamination neighborhood $\mathcal{P}_\epsilon = \{(1-\epsilon)P_0+\epsilon Q : Q \text{ arbitrary}\}$. The total-variation ambiguity set of Section 10.1 (with ϕ-divergence = TV distance) is precisely Huber's contamination neighbourhood – robust statistics' earliest ambiguity-set formulation predates the DRO literature by three decades. The connection is closer than terminological: Huber's M-estimators minimize a "robust loss" (the Huber function) that is, as shown in Section 10.5, the dual of the contamination-robust problem's Lagrangian – robust statistics' M-estimation and DRO's regularization-connection are the same mathematical structure viewed from two different disciplines.

Adversarial machine learning. Section 10.5's connection between Wasserstein-DRO and adversarial training (Madry et al., 2018) places the entire adversarial robustness literature within this volume's framework. The adversarial training objective $\min_\theta \frac{1}{N}\sum_i \max_{\|\delta\|_\infty \le \epsilon} \ell(\theta; x_i+\delta, y_i)$ is precisely the empirical version of Wasserstein-DRO with W_∞ geometry, making adversarial training a finite-sample stochastic approximation of distributionally robust empirical risk minimization – a connection that has generated substantial theoretical work in the last decade (Sinha et al., 2020; Staib and Jegelka, 2017) but whose practical implications (e.g. the right choice of ϵ and geometry in different deployment contexts) remain active research.

Financial risk management. The CVaR framework (Section 12.3) and the coherent-risk-measure axiomatics (Artzner et al., 1999) are the conceptual backbone of modern quantitative risk management, codified in regulatory frameworks (Basel III's Expected Shortfall requirement replaces VaR with CVaR at the 97.5% confidence level). The DRO perspective on CVaR – $\mathrm{CVaR}_\beta(Z) = \sup_{Q: D_{\mathrm{TV}}(Q,P) \le \beta/(1-\beta)} \mathbb{E}_Q[Z]$ (a TV-DRO representation) – gives a distributional-robustness interpretation to a measure that is often motivated purely axiomatically, suggesting that regulatory use of CVaR implicitly encodes a distributional-robustness assumption about tail behavior, an observation with implications for how stress testing and model validation should be understood.

Network resilience (Volume IV preview). The frameworks of this volume model uncertainty at a *single decision node*. Volume IV (Cascading Failures and Network Resilience) asks what happens when many such nodes interact: a robust decision at node v generates an uncertain input for node u, which in turn must be robust against it. The resulting *network of robust decisions* gives rise to qualitatively new phenomena – cascading failures, tipping points, and resilience thresholds – that cannot be captured by any single-node framework. Mathematically, the transition requires replacing the support function (a function on a normed space) by the spectral radius of an uncertainty-propagation operator (a function on a space of distributions on networks), a substantially harder object. Volume IV develops this theory from scratch, using the dual-norm and transport machinery of this volume as foundational

prerequisites.

Example — A concrete network-robustness problem motivating Volume IV

Consider a supply chain with $n = 5$ nodes; node i receives uncertain demand ξ_i and chooses inventory x_i. Robust feasibility requires $x_i \geq \xi_i$ for all $\xi_i \in \mathcal{U}_i$. But inventory at node i also generates uncertain supply for node j downstream: the uncertainty $\mathcal{U}_j$ that node j faces depends on node i's actual inventory choice x_i and its own demand realization. The "robust counterpart" for node j is thus an uncertain robust optimization problem (the uncertainty set $\mathcal{U}_j$ is itself uncertain, depending on x_i) – a two-level robustness structure not covered by any single-node framework. For $n = 5$ this nests to 5 levels; for general n and cyclic networks it creates a fixed-point problem (robust feasibility is self-referential). Volume IV's spectral-radius approach resolves this fixed-point structure.

13.4 A Reader's Map of the Literature

The four frameworks of this volume each have a primary reference treatise: Ben-Tal, El Ghaoui, and Nemirovski (2009) for RO (Part 9); Shapiro, Dentcheva, and Ruszczyński (2021) for SP (Part 12); Villani (2009) and Mohajerin Esfahani and Kuhn (2018) for Wasserstein-DRO (Part 10); and Ben-Haim (2006) for info-gap and Gilboa–Schmeidler (1989) for Choquet/ambiguity (Part 11). Boyd and Vandenberghe (2004) covers the convex-optimization prerequisites for all four, and Rockafellar (1970) provides the deepest common mathematical root.

For the connections between DRO and machine learning (Section 10.5 and Section 13.3), the key papers are Sinha, Namkoong, and Duchi (2020) on certifiable distributional robustness and Madry et al. (2018) on adversarial training. For robust statistics (Section 13.3), Huber (1981) remains the primary reference, complemented by the more recent treatment in Hampel et al. (1986). For financial risk and coherent measures, Artzner, Delbaen, Eber, and Heath (1999) established the axiomatic framework, and Föllmer and Schied (2011) give a comprehensive mathematical treatment.

Readers wishing to pursue the algorithmic side (Appendix C's interior-point and SDDP complexity) beyond what this volume provides should consult Nesterov and Nemirovski (1994) for interior-point methods and Shapiro, Tekaya, da Costa, and Pinto (2013) for SDDP implementation details and convergence results under Markovian uncertainty.

A note on notation across the literature. The field lacks a unified notation: "$\mathcal{U}$" is an uncertainty set in Part 9 but a "set of priors" in Part 11 (where Part 11 uses Γ); "λ" is a Lagrange multiplier dual to the Wasserstein-ball constraint in Part 10 but a Lipschitz constant elsewhere; "ϵ" is the Wasserstein radius in Section 10.2 but the approximation tolerance in Appendix C's complexity bounds. This appendix's glossary (Appendix D) standardises notation within this volume; readers moving to primary sources should expect these collisions and resolve them from context.

PART 13

Technical Appendices

Self-contained reference material on the convex-analytic, probabilistic, and computational tools used throughout Parts 9–12, collected here for ease of lookup.

A. Convex Analysis and Conic Duality Toolkit

Overview

This appendix collects, in one place, the convex-analytic facts (conjugate functions, support functions, conic duality, Hölder/dual-norm pairs) used repeatedly across Parts 9–12, together with proofs that were stated only sketchily in the main text. It is intended as a reference, not a first introduction – readers seeking motivation should consult Section 9.1 and Section 9.3 first.

Definition: Convex conjugate

For $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the *convex conjugate* (Legendre–Fenchel transform) is

$$f^*(\mathbf{y}) \triangleq \sup_{\mathbf{x}\in\mathbb{R}^n} (\mathbf{y}^\top \mathbf{x} - f(\mathbf{x})).$$

f^* is always convex (as a supremum of affine functions of $\mathbf{y}$), regardless of whether f itself is convex.

Connection to support functions. If $f = \delta_{\mathcal{U}}$, the indicator function of a set $\mathcal{U}$ ($\delta_{\mathcal{U}}(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{U}$, $+\infty$ otherwise), then $f^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathcal{U}} \mathbf{y}^\top \mathbf{x} = \sigma_{\mathcal{U}}(\mathbf{y})$ – the support function of $\mathcal{U}$ (Section 9.1) is exactly the conjugate of $\mathcal{U}$'s indicator function. This identifies the support-function machinery of Part 9 as a special case of the general conjugate-duality calculus, and explains why conjugate-duality identities (sums, scalings, compositions) translate directly into support-function identities such as the Minkowski-sum theorem of Section 9.1.

EXAMPLE — Conjugate of a quadratic

For $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x}$ with $\mathbf{Q} \succ 0$, $f^*(\mathbf{y}) = \sup_{\mathbf{x}}(\mathbf{y}^\top \mathbf{x} - \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x})$. Setting the gradient to zero, $\mathbf{y} = \mathbf{Q}\mathbf{x}^* \Rightarrow \mathbf{x}^* = \mathbf{Q}^{-1}\mathbf{y}$, giving $f^*(\mathbf{y}) = \mathbf{y}^\top \mathbf{Q}^{-1}\mathbf{y} - \frac{1}{2}\mathbf{y}^\top \mathbf{Q}^{-1}\mathbf{y} = \frac{1}{2}\mathbf{y}^\top \mathbf{Q}^{-1}\mathbf{y}$. For $\mathbf{Q} = \mathbf{I}$ (the case $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$), $f^*(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_2^2$ – the squared Euclidean norm is self-conjugate (up to the constant $\frac{1}{2}$), a fact used implicitly in Section 10.2's Wasserstein-DRO dual derivations whenever the cost function is quadratic.

Theorem: Fenchel–Young inequality

For any f and its conjugate f^*, $\mathbf{x}^\top \mathbf{y} \le f(\mathbf{x}) + f^*(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y}$, with equality iff $\mathbf{y} \in \partial f(\mathbf{x})$ (the subdifferential).

Proof. Immediate from the definition of f^* as a supremum: $f^*(\mathbf{y}) \ge \mathbf{y}^\top \mathbf{x} - f(\mathbf{x})$ for the particular $\mathbf{x}$ in question, rearranging to the stated inequality. Equality holds iff this particular $\mathbf{x}$ attains the supremum defining $f^*(\mathbf{y})$, which (for convex f) occurs precisely when $\mathbf{y}$ is a subgradient of f at $\mathbf{x}$. ■

Example — Fenchel–Young as a generalised Cauchy–Schwarz / Hölder inequality

For $f(\mathbf{x}) = \frac{1}{p}\|\mathbf{x}\|_p^p$ $(p > 1)$, $f^*(\mathbf{y}) = \frac{1}{p^*}\|\mathbf{y}\|_{p^*}^{p^*}$ $(\frac{1}{p} + \frac{1}{p^*} = 1)$. Fenchel–Young gives $\mathbf{x}^\top\mathbf{y} \le \frac{1}{p}\|\mathbf{x}\|_p^p + \frac{1}{p^*}\|\mathbf{y}\|_{p^*}^{p^*}$, Young's inequality for products – which, after the homogeneous rescaling $\mathbf{x} \to t\mathbf{x}$, $\mathbf{y} \to \mathbf{y}/t$ and optimizing over $t > 0$, recovers Hölder's inequality $\mathbf{x}^\top\mathbf{y} \le \|\mathbf{x}\|_p \|\mathbf{y}\|_{p^*}$ exactly. This is the same Hölder duality invoked, without proof, in Section 9.1's general norm-ball remarks – the conjugate-function machinery of this appendix is what makes that earlier claim rigorous.

Definition: Dual cone

For a cone $\mathcal{K} \subseteq \mathbb{R}^n$, the *dual cone* is $\mathcal{K}^* \triangleq \{\mathbf{y} : \mathbf{y}^\top\mathbf{x} \ge 0 \ \forall \mathbf{x} \in \mathcal{K}\}$. $\mathcal{K}$ is *self-dual* if $\mathcal{K}^* = \mathcal{K}$.

The three standard cones. The nonnegative orthant $\mathbb{R}^n_{\ge 0}$, the second-order (Lorentz) cone $\mathcal{Q}^n = \{(\mathbf{x}, t) : \|\mathbf{x}\|_2 \le t\}$, and the positive semidefinite cone $\mathcal{S}^n_{\succeq 0}$ are all self-dual – the algebraic fact underlying, respectively, LP duality, SOCP duality, and SDP duality. Conic optimization (Section 9.2's Remarks) generalises LP duality to any closed convex cone $\mathcal{K}$ exactly because $\mathcal{K}^*$ – whatever it is – plays the same structural role $\mathbb{R}^n_{\ge 0}$ plays in ordinary LP duality.

Example — Verifying self-duality of the SOC numerically

In $\mathbb{R}^3$, take $\mathcal{Q}^2 = \{(x_1, x_2, t) : \sqrt{x_1^2 + x_2^2} \le t\}$. For $\mathbf{a} = (3, 4, 5) \in \mathcal{Q}^2$ (since $\sqrt{9+16} = 5 \le 5$, boundary point) and a candidate dual element $\mathbf{b} = (3, 4, 5)$, check $\mathbf{a}^\top\mathbf{b} = 9 + 16 + 25 = 50 \ge 0$. ✓ For a generic interior point $\mathbf{a} = (1, 1, 2) \in \mathcal{Q}^2$ $(\sqrt{2} \approx 1.41 < 2)$ and $\mathbf{b} = (-1, -1, 2) \in \mathcal{Q}^2$ $(\sqrt{2} < 2)$, $\mathbf{a}^\top\mathbf{b} = -1 - 1 + 4 = 2 \ge 0$. ✓ In general, for $\mathbf{a} = (\mathbf{x}_a, t_a), \mathbf{b} = (\mathbf{x}_b, t_b) \in \mathcal{Q}^n$, $\mathbf{a}^\top\mathbf{b} = \mathbf{x}_a^\top\mathbf{x}_b + t_a t_b \ge -\|\mathbf{x}_a\|\,\|\mathbf{x}_b\| + t_a t_b \ge -t_a t_b + t_a t_b = 0$ by Cauchy–Schwarz and $\|\mathbf{x}_a\| \le t_a, \|\mathbf{x}_b\| \le t_b$ – confirming $\mathcal{Q}^n \subseteq (\mathcal{Q}^n)^*$, and the reverse inclusion follows by an analogous argument, giving self-duality.

Theorem: conic LP duality

For the primal $\min_{\mathbf{x}} \mathbf{c}^\top\mathbf{x}$ s.t. $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{x} \in \mathcal{K}$, the dual is $\max_{\mathbf{y}} \mathbf{b}^\top\mathbf{y}$ s.t. $\mathbf{c} - \mathbf{A}^\top\mathbf{y} \in \mathcal{K}^*$. Weak duality (primal $\ge$ dual) always holds; strong duality (equality, with both optima attained) holds under a constraint qualification (e.g. Slater's condition: existence of a strictly feasible $\mathbf{x}$ in the interior of $\mathcal{K}$).

Proof of weak duality. For feasible $\mathbf{x}, \mathbf{y}$: $\mathbf{c}^\top\mathbf{x} - \mathbf{b}^\top\mathbf{y} = \mathbf{c}^\top\mathbf{x} - (\mathbf{A}\mathbf{x})^\top\mathbf{y} = (\mathbf{c} - \mathbf{A}^\top\mathbf{y})^\top\mathbf{x} \ge 0$ since $\mathbf{c} - \mathbf{A}^\top\mathbf{y} \in \mathcal{K}^*$ and $\mathbf{x} \in \mathcal{K}$ (by definition of the dual cone). Hence $\mathbf{c}^\top\mathbf{x} \ge \mathbf{b}^\top\mathbf{y}$ for all feasible pairs, giving weak duality. ■

Example — A robust-LP duality gap closed by strong conic duality

Consider the robust counterpart of Section 9.3 reformulated as an SOCP (ellipsoidal uncertainty). Primal: $\min \mathbf{c}^\top\mathbf{x}$ s.t. $(\mathbf{x}, t) \in \mathcal{Q}^n$, $t \le b - \bar{\mathbf{a}}^\top\mathbf{x}$ (a thinly disguised SOC constraint). Because the ellipsoidal uncertainty set has nonempty interior and the problem is strictly feasible (Slater holds generically whenever $b - \bar{\mathbf{a}}^\top\mathbf{x} > 0$ is achievable with strict inequality for some $\mathbf{x}$), strong duality holds and the dual SOCP's optimal value exactly equals the primal robust-LP's optimal value – no duality gap, a fact implicitly used whenever Section 9.3's robust-counterpart reformulations are presented

as exact (rather than merely as upper or lower bounds).

Subdifferential calculus cheat-sheet. For f convex, $\partial f(\mathbf{x}) = \{\mathbf{g} : f(\mathbf{z}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{z}-\mathbf{x})\ \forall \mathbf{z}\}$. Key rules used throughout this volume: $\partial(\|\mathbf{x}\|_2) = \{\mathbf{x}/\|\mathbf{x}\|_2\}$ for $\mathbf{x} \neq \mathbf{0}$ (and the unit ball at $\mathbf{x} = \mathbf{0}$); $\partial(\|\mathbf{x}\|_1)_i = \operatorname{sgn}(x_i)$ for $x_i \neq 0$ (and $[-1,1]$ at $x_i = 0$); $\partial(\max_i f_i(\mathbf{x})) = \operatorname{conv}\{\partial f_i(\mathbf{x}) : i \in \arg\max_j f_j(\mathbf{x})\}$ (the convex hull of subgradients of the active maximizers) – this last rule is exactly what underlies the worst-case-direction characterizations (e.g. $u_i = \rho_i \operatorname{sgn}(x_i)$ in Section 9.1's support-function proof) used repeatedly across Parts 9–11.

Example — Subgradient of a pointwise maximum, numerically

$f(\mathbf{x}) = \max(x_1 + 2x_2,\ 3x_1 - x_2)$ at $\mathbf{x} = (1,1)$: the two pieces evaluate to $1+2 = 3$ and $3-1 = 2$; the first piece is active (strictly larger), so $\partial f(1,1) = \{(1,2)\}$, a singleton (since only one piece is active, f is differentiable here with gradient $(1,2)$). At $\mathbf{x} = (1, 0.5)$: pieces evaluate to $1+1 = 2$ and $3 - 0.5 = 2.5$; the second is active, $\partial f = \{(3,-1)\}$. At the boundary $x_1 + 2x_2 = 3x_1 - x_2 \Leftrightarrow x_2 = \frac{2}{3}x_1$, e.g. $\mathbf{x} = (3,2)$ (both pieces $= 7$): both pieces are active, and $\partial f(3,2) = \operatorname{conv}\{(1,2),(3,-1)\}$, the line segment between the two gradient vectors – a non-singleton subdifferential exactly at the kink, illustrating the max-rule's convex-hull clause concretely.

Historical notes. The conjugate-duality framework of this appendix is due to Fenchel (1949) and Rockafellar's systematic development in *Convex Analysis* (1970), already cited in Section 9.3's historical notes as the deepest common root of this volume's machinery. Conic duality's modern form, unifying LP/SOCP/SDP duality under the single dual-cone formalism presented here, crystallised somewhat later, in the interior-point-method literature of the 1990s (Nesterov and Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*, 1994) – the same period, and overlapping authorship, as the Ben-Tal–Nemirovski robust-optimization program discussed throughout Part 9's historical notes, which is no coincidence: robust optimization's reformulation toolkit *is*, in large part, an application of the slightly earlier conic-duality and interior-point machinery to a new class of (worst-case) problems.

B. Probabilistic Background: Concentration, Divergences, and Optimal Transport

Overview

This appendix collects the probabilistic prerequisites for Part 10 (DRO) and the probabilistic guarantee results of Section 9.4 and Section 12.3: concentration inequalities, ϕ-divergences, and the basic vocabulary of optimal transport. As with Appendix A, the goal is reference completeness rather than pedagogical introduction.

Theorem: Hoeffding's inequality

For independent bounded random variables $X_i \in [\alpha_i, \beta_i]$, $i = 1, \ldots, n$, and $S_n = \sum_i X_i$,

$$\Pr\left[S_n - \mathbb{E}[S_n] \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_i (\beta_i - \alpha_i)^2}\right).$$

Proof sketch. Via the Chernoff/Markov technique: $\Pr[S_n - \mathbb{E}S_n \geq t] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda(S_n - \mathbb{E}S_n)}]$ for

any $\lambda > 0$; independence gives $\mathbb{E}[e^{\lambda(S_n - \mathbb{E}S_n)}] = \prod_i \mathbb{E}[e^{\lambda(X_i - \mathbb{E}X_i)}]$, and Hoeffding's lemma bounds each factor by $\exp(\lambda^2(\beta_i - \alpha_i)^2/8)$. Optimizing the resulting bound $\exp(-\lambda t + \lambda^2 \sum_i (\beta_i - \alpha_i)^2/8)$ over λ (setting the derivative to zero, $\lambda^* = 4t/\sum_i (\beta_i - \alpha_i)^2$) gives the stated bound. ■

Connection to Section 9.4. The Bertsimas–Sim probabilistic guarantee's Chernoff-type bound $\exp(-\Gamma_i^2/2n_i)$ is precisely a Hoeffding-type tail bound applied to the sum $\sum_j z_{ij}$ of independent symmetric $[-1, 1]$ random variables ($\beta_i - \alpha_i = 2$ for each, giving the stated exponent up to the constant), making explicit that the entire budgeted-uncertainty probabilistic-guarantee apparatus of Section 9.4 is a direct instance of this appendix's Hoeffding inequality.

Example — Hoeffding bound, numerically, vs. the exact Binomial tail

For $n = 20$ i.i.d. symmetric ± 1 variables and $t = 8$ (i.e. $\Pr[S_{20} \geq 8]$), Hoeffding gives $\exp(-2 \cdot 64/(20 \cdot 4)) = \exp(-1.6) \approx 0.202$. The exact value (a Binomial tail, since $S_{20} \geq 8 \Leftrightarrow$ at least 14 of the 20 steps are $+1$) is $\Pr[\mathrm{Bin}(20, 0.5) \geq 14] \approx 0.0577$. The Hoeffding bound (0.202) is, as expected, conservative (looser) relative to the exact tail probability (0.058) by roughly a factor of 3.5 – a typical gap for moderate n, with the bound becoming relatively tighter (in a large-deviations/exponential-rate sense, though not necessarily in raw ratio) as $n \to \infty$.

Definition: ϕ-divergence

For a convex $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with $\phi(1) = 0$, the ϕ-divergence between distributions Q, P (with $Q \ll P$) is $D_\phi(Q\|P) = \int \phi\left(\frac{dQ}{dP}\right) dP$. Special cases: $\phi(t) = t \log t$ gives KL divergence; $\phi(t) = (t-1)^2$ gives the χ^2-divergence; $\phi(t) = \frac{1}{2}|t - 1|$ gives total variation.

Connection to Section 10.1. The ϕ-divergence balls $\{Q : D_\phi(Q\|\hat{P}_N) \leq \eta\}$ introduced informally in Section 10.1 are precisely sublevel sets of this appendix's D_ϕ; the choice of ϕ (KL vs. χ^2 vs. TV) determines both the ambiguity set's shape and the tractability of the resulting DRO reformulation (Section 10.3), with KL giving the cleanest exponential-family dual and χ^2 giving an exact SOCP/SDP-representable dual for many base losses.

Example — Comparing three ϕ-divergences for the same perturbation

Let $P = \mathrm{Bernoulli}(0.5)$ and $Q = \mathrm{Bernoulli}(0.6)$. KL: $D_{\mathrm{KL}}(Q\|P) = 0.6 \log \frac{0.6}{0.5} + 0.4 \log \frac{0.4}{0.5} = 0.6(0.182) + 0.4(-0.223) = 0.109 - 0.089 = 0.0202$. χ^2: $D_{\chi^2}(Q\|P) = \sum_x \frac{(Q(x)-P(x))^2}{P(x)} = \frac{0.01}{0.5} + \frac{0.01}{0.5} = 0.04$. TV: $D_{\mathrm{TV}}(Q\|P) = \frac{1}{2}(|0.6 - 0.5| + |0.4 - 0.5|) = 0.1$. All three correctly register $Q \neq P$, but on different numerical scales $(0.02$ vs. 0.04 vs. $0.1)$ – a reminder that the ambiguity-set radius η in Section 10.1's ϕ-divergence balls is *not* comparable across different choices of ϕ without an explicit conversion, a common source of confusion when comparing DRO results reported with different divergence choices in the literature.

Definition: Wasserstein distance (Kantorovich formulation)

For probability measures P, Q on a metric space (Ξ, d), the order-p Wasserstein dis-

tance is

$$W_p(P,Q) \triangleq \left(\inf_{\pi \in \Pi(P,Q)} \int d(\xi,\xi')^p \, d\pi(\xi,\xi') \right)^{1/p},$$

where $\Pi(P,Q)$ is the set of joint distributions (couplings) with marginals P and Q.

Connection to Section 10.2. The Wasserstein-DRO ambiguity set $\{Q : W_p(Q,\hat{P}_N) \le \epsilon\}$ of Section 10.2 is exactly the ϵ-ball in this appendix's W_p metric, and the dual reformulation result of Section 10.2/10.3 (via Kantorovich duality) is the optimization-theoretic payoff of the coupling-based definition given here: Kantorovich duality states $W_1(P,Q) = \sup_{\|f\|_{\text{Lip}} \le 1} (\mathbb{E}_P[f] - \mathbb{E}_Q[f])$, converting the (hard, infinite-dimensional) coupling infimum into a (tractable) supremum over 1-Lipschitz functions, the precise mechanism underlying every Wasserstein-DRO dual reformulation in Part 10.

Example — Computing W_1 between two discrete distributions explicitly

P puts mass $(0.5, 0.5)$ on $\{0,1\} \subset \mathbb{R}$; Q puts mass $(0.3, 0.7)$ on the same support. Any coupling π has marginals summing correctly; the optimal transport plan here is the "obvious" one: move 0.2 units of mass from P's point 0 to Q's point 1 (since Q has more mass at 1 and less at 0 than P), at cost $0.2 \times |1-0| = 0.2$, with the remaining 0.3 units at each point left in place (cost 0). Hence $W_1(P,Q) = 0.2$. This matches the general 1-D formula $W_1(P,Q) = \int_{\mathbb{R}} |F_P(x) - F_Q(x)| \, dx$ (with F the CDFs): here $F_P(x) = 0.5$ for $x \in [0,1)$ and $F_Q(x) = 0.3$ for $x \in [0,1)$, giving $|F_P - F_Q| = 0.2$ over an interval of length 1, integrating to 0.2. ✓

Theorem: Pinsker's inequality

$$D_{\text{TV}}(Q\|P) \le \sqrt{\tfrac{1}{2} D_{\text{KL}}(Q\|P)}.$$

Proof sketch. Follows from the variational representation of TV distance combined with a second-order Taylor bound on D_{KL} near $Q = P$; a full proof proceeds via the data-processing inequality applied to the binary experiment that best distinguishes P from Q, a standard but somewhat technical information-theoretic argument omitted here for brevity (see the references in this appendix's historical notes). ■

Example — Checking Pinsker's inequality against the worked KL/TV example above

From the ϕ-divergence example above, $D_{\text{KL}}(Q\|P) = 0.0202$ and $D_{\text{TV}}(Q\|P) = 0.1$ (for the same P, Q pair). Pinsker's bound gives $\sqrt{0.0202/2} = \sqrt{0.0101} \approx 0.1005$. Indeed $0.1 \le 0.1005$ – the inequality holds, and holds nearly *tightly* for this example (within 0.5%), illustrating that Pinsker's bound, while only an inequality in general, can be close to exact for small perturbations – consistent with its derivation via a local (Taylor) expansion around $Q = P$.

Historical notes. Hoeffding's inequality (1963, "Probability Inequalities for Sums of Bounded Random Variables") is part of the broader large-deviations/concentration-of-measure literature whose systematic development (Bernstein, Chernoff, Hoeffding, and later Talagrand's more general concentration inequalities of the 1990s) underlies essentially all finite-sample probabilistic guarantees in this volume, including Section 9.4's Bertsimas–Sim bound and Section 10.4's sample-complexity results. The Wasserstein distance, despite its prominence in modern DRO (Section 10.2), has 20th-century roots in the transportation problem of

Monge (1781) and its rigorous formalization by Kantorovich (1942, for which Kantorovich shared the 1975 Nobel Memorial Prize in Economics) – making it, like several tools in this volume, a 20th-century rediscovery and generalization of an 18th-century applied problem, mirroring the centuries-spanning intellectual lineages noted throughout this volume's historical notes.

C. Algorithms and Computational Complexity Reference

Overview

This appendix tabulates the computational complexity of the principal problem classes appearing throughout Parts 9–12, and sketches the standard solution algorithms referenced (but not derived in detail) in the main text.

Complexity classes used in this volume. P: polynomial-time solvable. LP, SOCP, SDP are all in P (via interior-point methods, polynomial in problem size and the logarithm of the desired accuracy). NP-hard problems (e.g. general mixed-integer robust optimization, or DRO with non-convex base loss) appear at several points in this volume and are flagged explicitly when relevant.

Theorem: interior-point complexity for conic programs

A self-concordant-barrier interior-point method solves an SDP of size $n \times n$ (matrix variable) and m constraints to accuracy ϵ in $O(\sqrt{n}\log(1/\epsilon))$ Newton iterations, each costing $O(n^3 + mn^2)$ arithmetic operations (for a dense Newton step) – giving an overall polynomial-time algorithm, though with a higher per-iteration cost than LP's simplex or interior-point methods (which scale as $O(n^{2.5})$ or better per iteration for comparable problem sizes).

Practical implication for Part 9. The asymptotic complexity gap between LP (Bertsimas–Sim's budgeted uncertainty, Section 9.4) and SDP (the most general conic robust counterparts, Section 9.2's Remarks) explains, in computational terms, why Section 9.4's historical notes describe the Bertsimas–Sim LP reformulation as having unlocked applications (national-scale unit commitment, with n in the hundreds of thousands) that were, in 2004, beyond reach for SDP-based robust formulations of comparable size – the $O(n^3)$ vs. effectively-linear per-iteration cost difference compounds dramatically at that scale.

Example — A worked complexity comparison

Solving a robust portfolio problem with $n = 1000$ assets: an LP-representable formulation (box or budgeted uncertainty) with simplex/interior-point costs roughly $O(n^{2.5}) \approx 3 \times 10^7$ operations per iteration, $O(\log n)$ to $O(\sqrt{n})$ iterations – seconds on modern hardware. An SDP-representable formulation (full covariance-ellipsoid robust counterpart with additional matrix constraints) costs $O(n^3) = 10^9$ operations per iteration – minutes to hours, and at $n = 10000$ the gap (LP: $\approx 10^{8.5}$; SDP: 10^{12} per iteration) becomes the difference between routinely solvable and effectively intractable on standard hardware, concretizing the abstract complexity-class distinction with realistic numbers.

Theorem: SDDP convergence rate

Under standard regularity conditions (bounded relatively complete recourse, finitely many stage-wise scenarios per node), SDDP's lower bound $\underline{V}_1^{(k)}$ converges to the true optimal value V_1^* almost surely as the iteration count $k \to \infty$, with the optimality gap $V_1^* - \underline{V}_1^{(k)}$ decreasing, empirically, at a rate often close to $O(1/\sqrt{k})$ in practice (no tight general worst-case rate is known; the algorithm's practical efficiency, noted in Section 12.4's historical notes, considerably outpaces its proven worst-case guarantees).

The gap between theory and SDDP practice. This appendix's SDDP convergence theorem

deliberately states only an asymptotic (eventual) guarantee, because no polynomial worst-case iteration bound is known for SDDP on general multi-stage problems – a striking contrast with the polynomial guarantees available for LP/SOCP/SDP (this appendix's interior-point theorem) and a partial explanation for why Section 12.5's checklist recommends SDDP only after confirming, via the extensive-form-size heuristic (b^{T-1}), that decomposition is actually necessary: SDDP's lack of a worst-case polynomial guarantee makes it a pragmatically excellent but theoretically incompletely-understood algorithm, an unusual combination for a method with three decades of large-scale industrial deployment (Section 12.4's historical notes).

Example — Tracking SDDP's empirical convergence on a small instance

For a $T = 4$-stage inventory problem with 5 scenarios per stage, SDDP's lower bound typically progresses as (illustrative, rounded): iteration 1: $\underline{V} = 120$; iteration 5: $\underline{V} = 145$; iteration 20: $\underline{V} = 158$; iteration 50: $\underline{V} = 159.8$; converged value $V^* \approx 160$. The gap shrinks from 40 (iter. 1) to 14.2 (5%, iter. 50) roughly consistent with an $O(1/\sqrt{k})$-type decay ($\sqrt{50/1} \approx 7.1$, and $40/14.2 \approx 2.8$ – not an exact match, illustrating that the "often close to $O(1/\sqrt{k})$" qualifier in this appendix's theorem is empirical and instance-dependent, not a precise theoretical rate).

Historical notes. The polynomial-time theory of interior-point methods for conic programming is due to Karmarkar's 1984 breakthrough for LP, generalised to general convex/conic programming by Nesterov and Nemirovski (1994, already cited in Appendix A's historical notes) via the self-concordant-barrier framework – the same theoretical apparatus that, applied specifically to the robust-counterpart reformulations of Part 9, gives the polynomial-time guarantees implicitly relied upon (but rarely proven explicitly) throughout this volume's tractability claims. SDDP's empirical-but-not-worst-case-polynomial convergence behaviour, discussed candidly in this appendix, has motivated a substantial line of more recent (2010s-2020s) algorithmic research seeking either improved worst-case guarantees under additional structural assumptions, or principled stopping rules – an active area not covered further in this volume.

D. Notation and Glossary

Overview

This appendix standardises the notation used throughout Parts 9–12 and provides brief glossary entries for the principal technical terms. Entries are grouped thematically to mirror the volume's four-part structure; within each group they appear in the order in which the concept first becomes important.

D.1 General Mathematical Notation

Sets and spaces. $\mathbb{R}^n$ denotes Euclidean n-space; $\mathbb{R}^n_{\geq 0}$ the nonnegative orthant. $\mathcal{S}^n$ is the space of $n \times n$ real symmetric matrices; $\mathcal{S}^n_{\succeq 0}$ its positive-semidefinite cone. $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector; $\mathbf{0}$ the zero vector (dimension clear from context). Calligraphic letters ($\mathcal{U}, \mathcal{K}, \mathcal{X}, \mathcal{P}$) denote sets; $|\mathcal{U}|$ the cardinality of a finite set. $\operatorname{int}(\mathcal{C})$ and $\operatorname{cl}(\mathcal{C})$ denote the interior and closure of a convex set $\mathcal{C}$.

Vectors and matrices. Boldface lowercase ($\mathbf{x}, \mathbf{a}, \boldsymbol{\xi}$) for vectors; boldface uppercase ($\mathbf{A}, \mathbf{Q}, \mathbf{P}$) for matrices. $\mathbf{x}^\top$ is the transpose of $\mathbf{x}$. $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$; $\mathbf{A}^{-1}$ its inverse (when square and non-singular); $\mathbf{A}^{-\top} = (\mathbf{A}^{-1})^\top$. $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive semidefinite; $\mathbf{A} \succ \mathbf{B}$ positive definite. $\operatorname{tr}(\mathbf{A}) = \sum_i A_{ii}$ is the trace; $\mathbf{A} \bullet \mathbf{B} = \operatorname{tr}(\mathbf{A}^\top \mathbf{B})$ the matrix inner product. $\operatorname{diag}(\mathbf{x})$ is the diagonal matrix with entries x_i; $\operatorname{vec}(\mathbf{A})$ the column-major vectorization.

Norms. $\|\mathbf{x}\|_p = (\sum_i |x_i|^p)^{1/p}$ for $p \geq 1$; $\|\mathbf{x}\|_\infty = \max_i |x_i|$; $\|\mathbf{x}\|_0 = |\{i : x_i \neq 0\}|$ (not a true norm). The dual norm of $\|\cdot\|$ is $\|\mathbf{y}\|_* = \sup_{\|\mathbf{x}\| \leq 1} \mathbf{y}^\top \mathbf{x}$; notably $\|\cdot\|_{p,*} = \|\cdot\|_{p^*}$ with $1/p + 1/p^* = 1$. $\|\mathbf{A}\|_F = \sqrt{\mathbf{A} \bullet \mathbf{A}}$ is the Frobenius norm; $\|\mathbf{A}\|_{\text{op}} = \sigma_{\max}(\mathbf{A})$ the operator (spectral) norm.

Functions. $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is the extended real-valued convention used throughout (avoids case distinctions on the domain). $\operatorname{dom}(f) = \{\mathbf{x} : f(\mathbf{x}) < +\infty\}$. $\partial f(\mathbf{x})$ is the subdifferential (set of subgradients) at $\mathbf{x}$; $\nabla f(\mathbf{x})$ the gradient when f is differentiable. $\delta_{\mathcal{C}}$ is the indicator of a convex set $\mathcal{C}$ (0 on $\mathcal{C}$, $+\infty$ off it). f^* is the convex conjugate (Legendre–Fenchel transform) of f. $[a]_+ = \max(a, 0)$ is the positive part.

Probability. P, Q, μ, ν denote probability measures; $\hat{P}_N = (1/N) \sum_{i=1}^N \delta_{\xi_i}$ the empirical measure on N samples $\{\xi_i\}$. $\mathbb{E}_P[f(\boldsymbol{\xi})]$ is the expectation of $f(\boldsymbol{\xi})$ under P; $\operatorname{Var}_P[Z]$ the variance. $Z \sim P$ means Z is distributed according to P. $\Pr[\mathcal{E}]$ is the probability of event $\mathcal{E}$. $\Pi(P, Q)$ is the set of couplings (joint distributions with marginals P and Q). F_P denotes the CDF of P.

D.2 Part 9 (Robust Optimization) Notation

Symbol	Meaning (first use)
$\mathcal{U}$	Uncertainty set (Section 9.1). Contains the realised parameter $\mathbf{a}$.
$\bar{\mathbf{a}}, \hat{\mathbf{a}}$	Nominal (centre) value of the uncertain parameter.
$\boldsymbol{\rho}, \Gamma$	Budget/scale parameter(s) for the uncertainty set (Sec. 9.1, 9.4).
$\sigma_{\mathcal{U}}(\mathbf{x})$	Support function of $\mathcal{U}$ at direction $\mathbf{x}$: $\sup_{\mathbf{a} \in \mathcal{U}} \mathbf{a}^\top \mathbf{x}$ (Sec. 9.1).
$\mathcal{K}, \mathcal{K}^*$	Cone and its dual cone (Sec. 9.2, App. A).
$\mathcal{Q}^n$	Second-order (Lorentz) cone in $\mathbb{R}^{n+1}$ (Sec. 9.2).

$\mathbf{P}_j$	Perturbation matrix in the general uncertainty parametrisation $\mathbf{a} = \bar{\mathbf{a}} + \sum_j \xi_j \mathbf{P}_j \mathbf{e}_j$ (Sec. 9.1).
z_j, z_j^+, z_j^-	Auxiliary variables in the Bertsimas–Sim LP reformulation (Sec. 9.4).
$\hat{\alpha}(\mathbf{x}, r_c)$	Info-gap robustness function (Sec. 11.5).
$\bar{\Pi}$, Π^*	Worst-case and nominal return vectors in the portfolio example (Sec. 9.4).

D.3 Part 10 (Distributionally Robust Optimization) Notation

Symbol	Meaning (first use)
$\mathcal{P}$	Ambiguity set: a set of probability measures (Sec. 10.1).
$D_\phi(Q\|P)$	ϕ-divergence from P to Q (Sec. 10.1, App. B).
$W_p(P, Q)$	Order-p Wasserstein distance (Sec. 10.2, App. B).
ϵ	Wasserstein-ball radius (Sec. 10.2).
π	Coupling (transportation plan) between two measures (Sec. 10.2, App. B).
$c(\xi, \xi')$	Ground transportation cost in the Wasserstein metric (Sec. 10.2).
κ	Lipschitz/growth rate parameter in Wasserstein-DRO duality (Sec. 10.2).
λ	Lagrange multiplier dual to the Wasserstein-ball constraint (Sec. 10.3).
$\hat{R}_N(\theta)$	Empirical risk on N samples at parameter θ (Sec. 10.4).
$\Omega(\theta)$	Regularisation term (e.g. $\|\theta\|_2^2$); connection to Wasserstein-DRO (Sec. 10.5).
δ^{adv}	Adversarial perturbation in adversarial training (Sec. 10.5).

D.4 Part 11 (Decision Under Deep Uncertainty) Notation

Symbol	Meaning (first use)
Γ	Set of priors (Sec. 11.1); distinct from budget parameter Γ of Sec. 9.4 (context disambiguates).
ν	Non-additive capacity / Choquet measure (Sec. 11.3).
$\bar{\nu}$	Conjugate capacity: $\bar{\nu}(A) = 1 - \nu(A^c)$ (Sec. 11.3).
λ	Hurwicz optimism coefficient $\lambda \in [0, 1]$ (Sec. 11.2).
$V_{\text{Ch}}(f, \nu)$	Choquet integral of act f against capacity ν (Sec. 11.3).
$\underline{E}_\Gamma[Z]$, $\bar{E}_\Gamma[Z]$	Lower/upper expectations over Γ (Sec. 11.4).
$R_{\text{minimax}}(\mathbf{x})$	Minimax regret at decision $\mathbf{x}$ (Sec. 11.2).
α	Info-gap horizon of uncertainty (Sec. 11.5).
r_c	Critical performance threshold in info-gap (Sec. 11.5).

D.5 Part 12 (Stochastic Programming) Notation

Symbol	Meaning (first use)
$Q(\mathbf{x}, \xi)$	Recourse function (second-stage value given first-stage $\mathbf{x}$ and scenario ξ) (Sec. 12.1).

$\mathcal{Q}(\mathbf{x})$	Expected recourse $\mathbb{E}_\xi[Q(\mathbf{x},\xi)]$ (Sec. 12.1).
c, h, p	Newsvendor costs: production, holding, shortage (Sec. 12.1).
$\mathbf{W}, \mathbf{T}, \mathbf{h}(\xi)$	Recourse matrix, technology matrix, rhs in two-stage SP (Sec. 12.1).
ξ^s, p^s	Scenario and its probability (Sec. 12.2).
$\text{VaR}_\beta(Z)$	Value-at-Risk at level β (Sec. 12.3).
$\text{CVaR}_\beta(Z)$	Conditional Value-at-Risk (expected shortfall) (Sec. 12.3).
$V_t(\mathbf{x})$	Stage-t value function in multi-stage SP (Sec. 12.4).
$\underline{V}_t^{(k)}$	SDDP lower-bound approximation after k iterations (Sec. 12.4, App. C).
$\mathbf{x}_t$, $\mathbf{y}_t$	Stage-t state and control variables (Sec. 12.4).
b, T	Branching factor and horizon of the scenario tree (Sec. 12.4).

D.6 Glossary of Key Terms

Adjustable Robust Optimization (ARO). An extension of static RO in which some decision variables (recourse variables) are allowed to be functions of the observed uncertainty, rather than fixed before uncertainty is revealed (Section 9.5).

Ambiguity set. A set $\mathcal{P}$ of probability distributions, used in DRO to model the case where the true distribution is unknown but known to lie within $\mathcal{P}$ (Part 10).

Benders cut. A supporting hyperplane of a (convex) value function $V_t(\mathbf{x})$, computed from dual variables of a stage-$(t+1)$ LP; Benders cuts are the building blocks of SDDP approximations (Section 12.4).

Capacity (Choquet). A monotone set function $\nu : 2^\Omega \to [0,1]$ with $\nu(\emptyset) = 0$, $\nu(\Omega) = 1$, not necessarily additive; generalises probability measures to model non-additive beliefs or risk assessments (Section 11.3).

Chance constraint. A constraint of the form $\Pr[\mathbf{a}(\boldsymbol{\xi})^\top \mathbf{x} \le b] \ge 1-\beta$, requiring the constraint to hold with probability at least $1-\beta$ over the uncertainty $\boldsymbol{\xi}$ (Section 12.3).

Conic program. An optimization problem of the form $\min \mathbf{c}^\top \mathbf{x}$ s.t. $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{x} \in \mathcal{K}$, where $\mathcal{K}$ is a closed convex cone; LP ($\mathcal{K} = \mathbb{R}^n_{\ge 0}$), SOCP ($\mathcal{K} = \mathcal{Q}^n$), and SDP ($\mathcal{K} = \mathcal{S}^n_{\succeq 0}$) are all conic programs (Section 9.2, Appendix A).

Conjugate function. For $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the function $f^*(\mathbf{y}) = \sup_{\mathbf{x}}(\mathbf{y}^\top \mathbf{x} - f(\mathbf{x}))$; always convex, and $f^{**} = f$ when f is convex and lower semicontinuous (Appendix A).

CVaR (Conditional Value-at-Risk). The expected value of the β-fraction worst outcomes of a random variable Z: $\text{CVaR}_\beta(Z) = \inf_t\{t + \frac{1}{1-\beta}\mathbb{E}[\max(Z-t,0)]\}$. Coherent risk measure. Equals VaR_β for continuous distributions (Section 12.3).

DRO (Distributionally Robust Optimization). A framework minimizing the worst-case expected cost over all distributions in an ambiguity set $\mathcal{P}$: $\min_{\mathbf{x}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\mathbf{x}, \boldsymbol{\xi})]$ (Part 10).

Info-gap (IGDT). A non-probabilistic robustness framework quantifying the maximum deviation α from a nominal model $\hat{\mathbf{u}}$ under which a decision $\mathbf{x}$ still satisfies a critical performance requirement r_c (Section 11.5).

Kantorovich duality. The dual representation $W_1(P,Q) = \sup_{\|f\|_{\text{Lip}} \le 1}(\mathbb{E}_P[f] - \mathbb{E}_Q[f])$, converting the coupling-based Wasserstein distance into a supremum over Lipschitz functions; key to tractable Wasserstein-DRO reformulations (Section 10.2, Appendix B).

Minimax regret. A decision criterion minimizing the maximum regret $\max_{\mathbf{a} \in \mathcal{U}}[\ell(\mathbf{x}, \mathbf{a}) -$

$\min_{\mathbf{x}'} \ell(\mathbf{x}', \mathbf{a})]$ over all scenarios; intermediate between worst-case (maximin) and expected-value criteria (Section 11.2).

Price of Robustness. The increase in (expected or nominal) cost incurred by adopting a robust solution rather than an expected-value or nominal-optimal solution; explicitly traded off against constraint-violation risk in the Bertsimas–Sim framework (Section 9.4).

Robust counterpart. The deterministic optimization problem obtained by replacing an uncertain constraint $f(\mathbf{x}, \mathbf{a}) \leq 0$ with its worst-case version $\max_{\mathbf{a} \in \mathcal{U}} f(\mathbf{x}, \mathbf{a}) \leq 0$; tractably reformulated via support functions and conic duality (Section 9.3).

SAA (Sample Average Approximation). A stochastic programming method replacing the true distribution P by the empirical measure $\hat{P}_N$ and solving the resulting finite scenario program; convergence guarantees under appropriate regularity conditions (Section 12.2).

SDDP (Stochastic Dual Dynamic Programming). A decomposition algorithm for multistage stochastic programs that approximates value functions by piecewise-linear lower envelopes (Benders cuts) and improves them via forward sampling and backward-pass dual computation (Section 12.4, Appendix C).

Support function. $\sigma_{\mathcal{U}}(\mathbf{x}) = \sup_{\mathbf{a} \in \mathcal{U}} \mathbf{a}^\top \mathbf{x}$; encodes the geometry of a convex set $\mathcal{U}$ and is the central object enabling tractable robust-counterpart reformulations via conic duality (Sections 9.1–9.3, Appendix A).

Wasserstein ball. The set $\{Q : W_p(Q, \hat{P}_N) \leq \epsilon\}$ of distributions within Wasserstein distance ϵ from the empirical measure; the most common ambiguity set in Wasserstein-DRO (Section 10.2).

E. Extended Proofs and Supplementary Derivations

Overview

This appendix supplies complete proofs or detailed derivations for results stated without full proof in the main text. Each entry identifies the section where the result was used, states it precisely, and provides the argument in self-contained form. The appendix is organised by Part rather than by order of appearance, grouping together results that share the same proof techniques.

E.1 Full Derivation of the Ellipsoidal Robust Counterpart (Section 9.1)

Let $\mathcal{U}_{\text{ell}} = \{\bar{\mathbf{a}} + \mathbf{P}\mathbf{u} : \|\mathbf{u}\|_2 \leq 1\}$ with $\mathbf{P} \in \mathbb{R}^{m\times k}$, and consider the constraint $\mathbf{a}^\top \mathbf{x} \leq b$ to be enforced for all $\mathbf{a} \in \mathcal{U}_{\text{ell}}$.

Theorem E.1 (Ellipsoidal robust counterpart)

The constraint $\mathbf{a}^\top \mathbf{x} \leq b \ \forall \mathbf{a} \in \mathcal{U}_{\text{ell}}$ is equivalent to the SOCP constraint

$$\bar{\mathbf{a}}^\top \mathbf{x} + \left\|\mathbf{P}^\top \mathbf{x}\right\|_2 \leq b.$$

Proof. The worst-case left-hand side is

$$\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) = \sup_{\|\mathbf{u}\|_2 \leq 1} (\bar{\mathbf{a}} + \mathbf{P}\mathbf{u})^\top \mathbf{x} = \bar{\mathbf{a}}^\top \mathbf{x} + \sup_{\|\mathbf{u}\|_2 \leq 1} \mathbf{u}^\top (\mathbf{P}^\top \mathbf{x}).$$

The inner supremum $\sup_{\|\mathbf{u}\|_2 \leq 1} \mathbf{u}^\top \mathbf{v}$ equals $\|\mathbf{v}\|_2$ (attained at $\mathbf{u}^* = \mathbf{v}/\|\mathbf{v}\|_2$ for $\mathbf{v} \neq \mathbf{0}$, and 0 for $\mathbf{v} = \mathbf{0}$), so

$$\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) = \bar{\mathbf{a}}^\top \mathbf{x} + \left\|\mathbf{P}^\top \mathbf{x}\right\|_2 .$$

The robust constraint $\sigma_{\mathcal{U}_{\text{ell}}}(\mathbf{x}) \leq b$ is therefore $\bar{\mathbf{a}}^\top \mathbf{x} + \left\|\mathbf{P}^\top \mathbf{x}\right\|_2 \leq b$, which is a second-order cone constraint in $(\mathbf{x}, b)$ — recognisable as $(\mathbf{P}^\top \mathbf{x}, b - \bar{\mathbf{a}}^\top \mathbf{x}) \in \mathcal{Q}^k$ — directly amenable to SOCP solvers. ■

Example — Worked numerical verification

Let $m = 2$, $\bar{\mathbf{a}} = (1, 2)$, $\mathbf{P} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.3 \end{pmatrix}$, $\mathbf{x} = (1, 1)$, $b = 3$. Then $\bar{\mathbf{a}}^\top \mathbf{x} = 3$ and $\mathbf{P}^\top \mathbf{x} = (0.5, 0.3)$, so $\left\|\mathbf{P}^\top \mathbf{x}\right\|_2 = \sqrt{0.25 + 0.09} = \sqrt{0.34} \approx 0.583$. The SOC constraint requires $3 + 0.583 = 3.583 \leq 3$, which fails – $\mathbf{x} = (1, 1)$ is *not* in the robust feasible set. Setting $b = 4$: $3.583 \leq 4$ holds. The worst-case realisation of $\mathbf{a}$ is $\bar{\mathbf{a}} + \mathbf{P}\mathbf{u}^*$ where $\mathbf{u}^* = \mathbf{P}^\top \mathbf{x}/\left\|\mathbf{P}^\top \mathbf{x}\right\|_2 = (0.5, 0.3)/0.583 \approx (0.858, 0.515)$: $\mathbf{a}^* \approx (1 + 0.5 \cdot 0.858, 2 + 0.3 \cdot 0.515) = (1.429, 2.155)$, and indeed $\mathbf{a}^{*\top} \mathbf{x} \approx 1.429 + 2.155 = 3.584 \approx 3 + \left\|\mathbf{P}^\top \mathbf{x}\right\|_2$. ✓

E.2 Dual Derivation for Wasserstein-DRO with Linear Loss (Section 10.2)

Let P be a distribution on $\Xi \subseteq \mathbb{R}^d$, $\hat{P}_N = (1/N) \sum_i \delta_{\xi_i}$ the empirical measure, and $\ell(\mathbf{x}, \xi) = \mathbf{x}^\top \xi$ (linear loss). The Wasserstein-DRO problem is

$$\sup_{Q:\, W_1(Q, \hat{P}_N) \leq \epsilon} \mathbb{E}_Q[\mathbf{x}^\top \boldsymbol{\xi}].$$

Theorem E.2 (Wasserstein-DRO dual for linear loss)

Under a 1-Lipschitz-bounded perturbation ($\|\mathbf{x}\|_* \le \kappa$), the supremum equals

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^\top \xi_i + \epsilon \|\mathbf{x}\|_* .$$

Proof. By Kantorovich duality,

$$W_1(Q, \hat{P}_N) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} (\mathbb{E}_Q[f(\boldsymbol{\xi})] - \mathbb{E}_{\hat{P}_N}[f(\boldsymbol{\xi})]),$$

so the constraint $W_1(Q, \hat{P}_N) \le \epsilon$ is equivalent to: for all f with Lipschitz constant ≤ 1, $\mathbb{E}_Q[f] \le \mathbb{E}_{\hat{P}_N}[f] + \epsilon$. The Lagrangian relaxation of the primal problem (perturbing the empirical samples by adversarial shifts $\delta_i \in \mathbb{R}^d$) gives

$$\sup_{Q:\, W_1(Q,\hat{P}_N)\le\epsilon} \mathbb{E}_Q[\mathbf{x}^\top \boldsymbol{\xi}] = \frac{1}{N}\sum_i \sup_{\delta_i:\, \|\delta_i\|\le\epsilon} \mathbf{x}^\top(\xi_i + \delta_i) = \frac{1}{N}\sum_i \left(\mathbf{x}^\top \xi_i + \epsilon \|\mathbf{x}\|_*\right),$$

where the inner supremum $\sup_{\|\delta\|\le\epsilon} \mathbf{x}^\top \delta = \epsilon \|\mathbf{x}\|_*$ uses the definition of the dual norm. Summing over i gives the stated result. ■

Implication. For ℓ_2 geometry ($\|\cdot\| = \|\cdot\|_2$, so $\|\cdot\|_* = \|\cdot\|_2$ as well), the worst-case expected loss is the empirical mean plus an ℓ_2-norm regularisation penalty, immediately yielding the connection to Tikhonov/ridge regularisation discussed in Section 10.5: the Wasserstein-DRO problem with linear loss is precisely the same as empirical risk minimization with an ℓ_2 penalty of magnitude ϵ on the model parameters.

E.3 CVaR as a Linear Program (Section 12.3)

The representation of CVaR as an LP, central to Section 12.3's tractability claims, is derived in full here.

Theorem E.3 (Rockafellar–Uryasev representation)

For a random loss $Z(\mathbf{x}, \boldsymbol{\xi})$ with finite expectation,

$$\mathrm{CVaR}_\beta(Z(\mathbf{x}, \cdot)) = \min_{t\in\mathbb{R}} \left\{ t + \frac{1}{1-\beta}\mathbb{E}[\max(Z(\mathbf{x}, \boldsymbol{\xi}) - t, 0)] \right\},$$

and the minimum is attained at $t^* = \mathrm{VaR}_\beta(Z(\mathbf{x}, \cdot))$.

Proof. Let $F_Z(z) = \Pr[Z \le z]$ be the CDF of $Z = Z(\mathbf{x}, \boldsymbol{\xi})$. By definition,

$$\mathrm{VaR}_\beta(Z) = \inf\{z : F_Z(z) \ge \beta\}, \quad \mathrm{CVaR}_\beta(Z) = \frac{1}{1-\beta}\int_\beta^1 \mathrm{VaR}_u(Z)\, du = \mathbb{E}[Z \mid Z \ge \mathrm{VaR}_\beta(Z)].$$

Define $g(t) = t + \frac{1}{1-\beta}\mathbb{E}[\max(Z - t, 0)]$. Then

$$\frac{dg}{dt} = 1 - \frac{1}{1-\beta}\Pr[Z > t] = 1 - \frac{1 - F_Z(t)}{1-\beta}.$$

This derivative is ≤ 0 for $t \le \mathrm{VaR}_\beta(Z)$ (since $1 - F_Z(t) \ge 1 - \beta$ for t below the β-quantile) and ≥ 0 for $t \ge \mathrm{VaR}_\beta(Z)$, so g is convex and minimised at $t^* = \mathrm{VaR}_\beta$. Substituting:

$$\begin{aligned} g(t^*) &= t^* + \frac{1}{1-\beta}\mathbb{E}[\max(Z - t^*, 0)] \\ &= t^* + \frac{1}{1-\beta}\mathbb{E}[(Z - t^*)\mathbf{1}_{Z>t^*}] \\ &= t^* + \frac{1}{1-\beta}\left(\mathbb{E}[Z\mathbf{1}_{Z>t^*}] - t^*\Pr[Z > t^*]\right) \\ &= t^*(1 - \Pr[Z > t^*]/(1-\beta)) + \frac{1}{1-\beta}\mathbb{E}[Z\mathbf{1}_{Z>t^*}]. \end{aligned}$$

For a continuous F_Z, $\Pr[Z > t^*] = 1 - \beta$ exactly, so the first term is $t^* \cdot 0 = 0$ and $g(t^*) = (1/(1-\beta))\mathbb{E}[Z\mathbf{1}_{Z>t^*}] = \mathrm{CVaR}_\beta(Z)$. For general (possibly discrete) F_Z, the argument carries through with appropriate treatment of the atom at t^*, using the standard convention for the conditional tail expectation. ■

LP formulation for finite scenarios. With N equally-weighted scenarios $\xi^1, \dots, \xi^N$ (probability $1/N$ each), the Rockafellar–Uryasev representation gives

$$\mathrm{CVaR}_\beta(Z(\mathbf{x}, \cdot)) \approx \min_{t, \mathbf{s}} \left\{ t + \frac{1}{N(1-\beta)} \sum_{i=1}^{N} s_i : s_i \ge Z(\mathbf{x}, \xi^i) - t,\ s_i \ge 0 \right\},$$

which is a linear program in $(t, \mathbf{s}) \in \mathbb{R} \times \mathbb{R}^N$ whenever $Z(\mathbf{x}, \xi^i)$ is linear in $\mathbf{x}$ — directly implementable in any LP solver and the basis for all of Section 12.3's CVaR-constrained reformulations.

E.4 The Choquet Integral: Definition and Comonotonic-Additivity (Section 11.3)

Definition E.4 (Choquet integral)

For a bounded measurable function $f : \Omega \to \mathbb{R}$ and a capacity ν, the Choquet integral is

$$\int f \, d\nu \triangleq \int_0^\infty \nu(\{f > t\})\, dt + \int_{-\infty}^0 (\nu(\{f > t\}) - 1)\, dt,$$

where the integrals on the right are ordinary Lebesgue integrals in t.

This "horizontal slicing" formula reduces to the standard Lebesgue integral when ν is a probability measure (since $\nu(\{f > t\}) = 1 - F_f(t)$ gives the standard formula $\mathbb{E}[f] = \int_0^\infty (1 - F_f(t))\, dt - \int_{-\infty}^0 F_f(t)\, dt$), confirming that Choquet integration generalises probability-weighted expectation.

Theorem E.5 (Comonotonic additivity)

If f and g are comonotone (there exists a measurable ordering such that $f(\omega_1) \ge f(\omega_2) \Leftrightarrow g(\omega_1) \ge g(\omega_2)$ for ν-a.e. ω_1, ω_2), then

$$\int (f + g)\, d\nu = \int f \, d\nu + \int g \, d\nu.$$

Proof. Comonotonicity means $f + g$ and f (resp. g) have the same level sets: $\{f + g > t\}$ can be written as a union of sets $\{f > s\} \cap \{g > t - s\}$ that coincide with $\{f > s\}$ and $\{g > t - s\}$ individually (since the two functions cannot simultaneously be "too low" when

they are comonotone). Formally, for comonotone f, g: $(f+g)^{-1}(t,\infty) = f^{-1}(s,\infty)$ for the appropriate $s = s(t)$, and the horizontal-slicing integral of $f+g$ separates into the sum of the horizontal-slicing integrals of f and g independently. ■

Non-additivity for non-comonotone functions. When f and g are counter-monotone (one high whenever the other is low — a "diversified" pair), the Choquet integral with a supermodular ν satisfies $\int (f+g)\,d\nu \le \int f\,d\nu + \int g\,d\nu$ (subadditivity), capturing the risk-reducing effect of diversification; for submodular ν the inequality reverses. This is the formal sense in which Choquet integration respects the non-additive structure of ν, discussed informally in Section 11.3.

E.5 Worst-Case Distribution in a Moment-Based Ambiguity Set (Section 10.1)

Let $\mathcal{P} = \{Q : \mathbb{E}_Q[\boldsymbol{\xi}] = \bar{\boldsymbol{\xi}},\ \mathrm{Cov}_Q(\boldsymbol{\xi}) = \square\}$ (moment-based ambiguity set) and $\ell(\mathbf{x}, \xi) = (\mathbf{a}^\top \xi - b)^+$ (hinge loss). A classical result characterises the worst-case distribution.

Theorem E.6 (Worst-case distribution is two-point)

Under the moment-based ambiguity set above, the worst-case expected hinge loss

$$\sup_{Q \in \mathcal{P}} \mathbb{E}_Q[(\mathbf{a}^\top \boldsymbol{\xi} - b)^+]$$

is attained by a distribution supported on at most two points.

Proof sketch. By the Richter–Rogosinski–Mulholland–Rogers moment theorem (and its modern extensions due to Isii, 1960), the supremum of a linear functional over a moment-constrained set of probability measures is attained at an extreme point of the feasible set in the weak-$*$ topology. For two first-and-second-moment constraints (mean and covariance) in dimension d, the Richter–Isii theorem guarantees that extreme points are supported on at most $d+1$ points. In the scalar case ($d = 1$, i.e. the loss depends only on the scalar $\mathbf{a}^\top \boldsymbol{\xi}$), extreme measures are supported on at most $2 = d+1$ points. The worst-case distribution explicitly concentrates mass p at $\xi_H = (b + \mathbf{a}^\top \bar{\boldsymbol{\xi}} + \sqrt{(1-p)\mathbf{a}^\top \square \mathbf{a}/p})/\mathbf{a}^\top \mathbf{1}$ and mass $1-p$ at a lower point ξ_L, with p chosen to match the variance constraint. ■

Closed-form worst-case bound. In the scalar version, the resulting worst-case expected hinge loss equals

$$\frac{1}{2}\left(\mu - b + \sqrt{(\mu - b)^2 + \sigma^2}\right),$$

where $\mu = \mathbb{E}[\xi]$ and $\sigma^2 = \mathrm{Var}[\xi]$ — a closed-form "robust newsvendor" formula that appears (without derivation) in Section 10.1's discussion of moment-based DRO and that reduces, in the symmetric case $\mu = b$, to $\sigma/2$.

E.6 Strong Duality for the Wasserstein-DRO Problem (Sketch, Section 10.3)

The central duality theorem of Section 10.3 (converting the $\sup_Q$ over the Wasserstein ball into a tractable inf over Lagrange multipliers) rests on an infinite-dimensional minimax interchange. A complete proof is beyond this volume's scope, but the key steps are:

Step 1 (*Weak duality*). For any $\lambda \ge 0$ and any 1-Lipschitz function f,

$$\mathbb{E}_Q[\ell(\mathbf{x}, \boldsymbol{\xi})] \le \mathbb{E}_Q[\ell(\mathbf{x}, \boldsymbol{\xi}) + \lambda\, d(\boldsymbol{\xi}, \boldsymbol{\xi}')] - \lambda \mathbb{E}_Q[d(\boldsymbol{\xi}, \boldsymbol{\xi}')] \le \mathbb{E}_{\hat{P}_N}[\ell(\mathbf{x}, \cdot) + \lambda \epsilon],$$

so the dual objective upper-bounds any primal value. This establishes $\sup_Q \mathbb{E}_Q[\ell] \le \inf_\lambda(\cdot)$.

Step 2 (*Attainment and strong duality*). Under mild regularity (Lipschitz ℓ, compact Ξ, or appropriate growth conditions), the Fenchel–Moreau minimax theorem applies to the bilinear Lagrangian, guaranteeing that strong duality holds and the $\inf_\lambda \sup_Q$ equals the $\sup_Q$ — i.e. there is no duality gap. This is the content of Theorem 3.4 in Mohajerin Esfahani and Kuhn (2018), whose proof uses the theory of upper-semicontinuous functions and Prokhorov compactness for probability measures on Polish spaces.

Step 3 (*Finite reformulation*). Once strong duality is established, substituting the worst-case $\xi^*(\xi_i, \lambda)$ (the adversarial perturbation of sample i) into the dual gives the finite-dimensional reformulation cited in Section 10.3. For linear loss, this yields the result of Theorem E.2; for general convex loss with Lipschitz constant κ, the reformulation is $\inf_{\lambda \geq 0}\{\lambda\epsilon + \frac{1}{N}\sum_i \sup_{\xi'}[\ell(\mathbf{x}, \xi') - \lambda d(\xi', \xi_i)]\}$.

E.7 Proof of the Bertsimas–Sim Probabilistic Guarantee (Section 9.4)

Section 9.4 states that, when constraint coefficients $a_{ij}(\boldsymbol{\xi})$ take independent symmetric values in $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$, the Bertsimas–Sim solution with budget Γ_i satisfies $\Pr[\text{constraint } i \text{ violated}] \leq \exp(-\Gamma_i^2/(2n_i))$. The derivation proceeds as follows.

Setup. Fix constraint i; drop subscripts. The constraint is $\sum_j a_j(\xi_j)x_j \leq b$, with $a_j = \bar{a}_j + \xi_j \hat{a}_j$, $\xi_j \in [-1, 1]$ i.i.d. symmetric. Violation occurs iff $\sum_j \xi_j \hat{a}_j x_j > b - \sum_j \bar{a}_j x_j$. Define $Z_j = \xi_j \hat{a}_j x_j$ (independent, mean zero, bounded in $[-\hat{a}_j|x_j|, \hat{a}_j|x_j|]$). The Bertsimas–Sim solution satisfies $\bar{\mathbf{a}}^\top \mathbf{x} + \Gamma \cdot (\text{optimal dual value}) \leq b$ (i.e. the nominal + protection term $\leq b$), so the residual "safe margin" is $b - \bar{\mathbf{a}}^\top \mathbf{x} \geq \Gamma \cdot \max_j \hat{a}_j|x_j|$.

Applying Hoeffding. By Hoeffding's inequality (Appendix B) with $\sum_j (2\hat{a}_j|x_j|)^2$ in the denominator:

$$\Pr\left[\sum_j Z_j > t\right] \leq \exp\left(\frac{-2t^2}{\sum_j (2\hat{a}_j|x_j|)^2}\right).$$

Set $t = \Gamma \max_j \hat{a}_j|x_j|$ and $\sum_j (2\hat{a}_j|x_j|)^2 \leq n \cdot (2\max_j \hat{a}_j|x_j|)^2$ (crudely bounding by the maximum term, $n =$ number of uncertain coefficients). Then

$$\Pr[\text{violation}] \leq \exp\left(\frac{-2\Gamma^2(\max_j \hat{a}_j|x_j|)^2}{n \cdot 4(\max_j \hat{a}_j|x_j|)^2}\right) = \exp\left(\frac{-\Gamma^2}{2n}\right),$$

the stated bound. ■ The bound is conservative (the crude maximization overstates $\sum_j (2\hat{a}_j|x_j|)^2$); tighter bounds using $\sum_j$ rather than $n \cdot \max$ appear in the literature but yield less-clean closed forms.

Bibliography

References

[1] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, NJ, 2009.

[2] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[3] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

[4] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[5] D. Bertsimas and M. Sim. Robust discrete optimization and network flows. *Mathematical Programming*, 98(1-3):49–71, 2003.

[6] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[7] A. Ben-Tal and A. Nemirovski. Robust optimization – methodology and applications. *Mathematical Programming*, 92(3):453–480, 2002.

[8] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

[9] Y. Goldberg and A. D. Ioffe. Conic programming and duality. *Mathematical Programming*, 72(1):1–29, 1996.

[10] A. L. Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21(5):1154–1157, 1973.

[11] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[12] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: a convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

[13] R. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[14] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[15] R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[16] B. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[17] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[18] C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

[19] L. V. Kantorovich. On the translocation of masses (in Russian). *Doklady Akademii Nauk SSSR*, 37(7-8):227–229, 1942.

[20] G. A. Hanasusanto and D. Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research*, 66(3):849–869, 2018.

[21] D. Ellsberg. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.

[22] I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

[23] D. Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.

[24] Y. Ben-Haim. *Info-Gap Decision Theory: Decisions Under Severe Uncertainty*. 2nd ed. Academic Press, Oxford, 2006.

[25] L. Hurwicz. Optimality criteria for decision making under ignorance. *Cowles Commission Discussion Paper: Statistics*, No. 370, 1951.

[26] J. Milnor. Games against nature. In *Decision Processes*, R. M. Thrall et al. (eds.), Wiley, 1954.

[27] A. Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

[28] F. H. Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin, Boston, 1921.

[29] L. J. Savage. *The Foundations of Statistics*. Wiley, New York, 1954.

[30] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. 2nd ed. Springer, New York, 2011.

[31] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. 3rd ed. SIAM, Philadelphia, 2021.

[32] G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3-4):197–206, 1955.

[33] M. V. F. Pereira and L. M. V. G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1-3):359–375, 1991.

[34] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

[35] A. Prékopa. *Stochastic Programming*. Kluwer, Dordrecht, 1995.

[36] H. Scarf. A min-max solution of an inventory problem. In K. Arrow, S. Karlin, and H. Scarf (eds.), *Studies in the Mathematical Theory of Inventory and Production*, Stanford University Press, 1958.

[37] A. J. King and R. T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1):148–162, 1993.

[38] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[39] W. Fenchel. On conjugate convex functions. *Canadian Journal of Mathematics*, 1(1):73–77, 1949.

[40] Y. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, 1994.

[41] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[42] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, 1964.

[43] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.

[44] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[45] K. Isii. The extrema of probability determined by generalized moments (I): bounded random variables. *Annals of the Institute of Statistical Mathematics*, 12(2):119–134, 1960.